University of Zurich[UZH]

Designing the Right Experiment Right.

Interactive Systems to Support Trade-off and Sample Size Decisions in HCI Experiment Design

Dissertation submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wissenschaften, Dr. sc.
(corresponds to Doctor of Science, PhD)

presented by
Alexander Eiselmayer
from Heilbronn, Germany

approved in Feburary 2023

at the request of

Prof. Dr. Chat Wacharamanotham
Prof. Dr. Alan Dix
Dr. Wendy E. Mackay

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, February 15, 2023

The Chairperson of the Doctoral Board: Prof. Elaine Huang, Ph.D.

Contents

List of Figures

List of Tables

Abstract

Controlled experiments enable researchers to empirically confirm causal relationships between a cause and its effect. In the field of Human–Computer Interaction, controlled experiments are frequently utilized to validate new artifacts and interaction techniques by comparing them to existing benchmarks. Throughout the process of designing experiments, researchers must make critical decisions that determine the success of the study, such as determining whether the findings indicate improvement or not. These decisions involve uncertainties in estimating parameters and weighing trade-offs.

This thesis explores ways to support researchers during the experimental design process. The three main contributions of this thesis include (1) *Touchstone2*—a web application that allows researchers to design and compare experimental designs and their trade-offs, which was evaluated in two workshops; (2) *Argus*—a web application that facilitates sample size planning; and (3) SPEED—a protocol for flexible sample size planning, supported by an application called SPEEDX. The work concludes by discussing the limitations of the research and highlighting potential opportunities for future exploration.

Acknowledgments

Ever since my teenage years, I have been interested in pursuing a Ph.D., even though I did not fully understand what it entailed. In the first week of my Master's studies, I coincidentally met Chat at an introductory event at the university, and I immediately expressed my interest to him. Several months later, I joined Chat's class on quantitative methods in Human–Computer Interaction (HCI). As the semester came to a close, Chat sent me the following email, which would mark the beginning of my Ph.D. journey:

> "*Hi Alex,*
> *Do you already have a plan for the last week of June and the first two weeks of July? There's a possibility for a research collaboration with a lab in Paris that you may be interested in.*
> *Cheers,*
> *Chat*"

Thrilled by this opportunity, I traveled to Paris to meet Wendy and Michel. After two weeks of incredible and inspiring experiences, working on projects and meeting amazing people, I began my master's thesis, which would serve as the foundation for my Ph.D.

Chat, I am tremendously grateful for that email and the doors it opened for me. Your guidance and mentorship have had a profound impact on me, both as a researcher and as an individual. Through challenging times and moments of celebration, I have learned and grown under your tutelage. Thank you, Chat, for accompanying me on this journey!

Wendy and Michel, you introduced me to the captivating world of academic research. While it may not have been officially possible, you have become co-advisors for my thesis, shaping my work, skills, and future. My time at Ex)Situ was genuinely inspiring, and it ignited my passion for research. The lessons I learned during the Bootcamp will continue to resonate long after my Ph.D. Thank you, Wendy and Michel, for inviting me and offering your unwavering support throughout my Ph.D.

Xiaoyi, Kasper, and Joanna, our collaborations on various projects have been invaluable. I deeply appreciate your professional guidance, effective teamwork, and delightful conversations. Thank you for enriching my work.

My enthusiasm for HCI research blossomed at the Ex)Situ lab. Alongside Wendy and Michel, I had the pleasure of interacting with exceptional PostDocs, Ph.D. candidates, and Master's students. Thank you Carla, Germán, Ignacio, Janin, Jean-Philippe, Liz, Michael, Miguel, Nicolas, Sally, Téo, and Viktor.

Although it all began in Paris, I couldn't have completed my Ph.D. without the unwavering support of the wonderful members of the ZPAC lab. Thank you Anton, Elaine, Lu, Natasha, and Nimra for supporting my journey until the end.

Lars, Suzanne, and Marine, without your assistance, I would not have reached this milestone. I am deeply grateful for your help with my Ph.D. proposal and this thesis! Beyond our work together, we have shared challenging moments and unforgettable fun, which has cemented our friendship. Thank you for your support, and I eagerly anticipate the adventures that lie ahead.

Abby, thank you for being my partner in crime, providing assistance with the Ph.D. and at home. Your passion for science, HCI, and research has been a constant source of motivation and focus for me. It's time for us to embark on our next adventure together.

Above all, I want to express my deepest gratitude to my parents, Susanne and Klaus, who have nurtured and shaped me into the person I am today. Without your unwavering support and encouragement, I would not be where I am now! Finally, I want to thank my brother Michael, who always lent me a listening ear and provided invaluable advice, despite having far more exciting work stories to share.

Onwards to the next challenge,
Alex

Chapter 1

Introduction

"If the design of an experiment is faulty,
any method of interpretation which makes it out to be decisive
must be faulty too."

—Ronald A. Fisher

Controlled experiments are used to establish cause–effect relationships and validate hypotheses in research. An illustrative example of this is the story recounted by Salsburg [2001] in his book *The Lady Tasting Tea*. Muriel Bristol claimed to be able to distinguish whether milk or tea was added first to a cup, and Ronald A. Fisher proposed a plan to empirically test her claim. Fisher's plan introduced important concepts such as the null hypothesis, randomization of conditions, statistical power, and sample size planning, all of which are crucial considerations in experimental decision-making.

Inspired by this event, Fisher described the process of designing and analyzing controlled experiments in his book *The Design of Experiments* [Fisher, 1937], which laid the groundwork for modern statistical methods. Since then, further developments and discussions within the scientific community have led to the establishment of guidelines and processes for researchers, including the reporting of analysis results. However, the core principle articulated by Fisher still holds true: if the experimental design is flawed, the credibility of the analysis results is called into question. Therefore, it is vital for researchers to meticulously design experiments and conduct thorough analyses to provide robust evidence either supporting or refuting their hypotheses.

The significance of careful experimental design is particularly evident in fields like Medicine. For instance, Polack et al. [2020] were entrusted with the task of assessing the safety and efficacy of the COVID-19 vaccine BNT162b2 (BioNTech-Pfizer vaccine). A total of 43,448 participants were recruited, receiving either the vaccine or a placebo. Given the profound impact of the pandemic on our lives over the past two years, it is of utmost importance to have reliable and trustworthy results from such studies.

Various scientific fields, including Psychology, Cognitive Science, Neuroscience, and Computer Science, derive significant benefits from employing sound experimental practice. These practices extend beyond Medicine and encompass domains involving humans, animals, or bacteria. Within Computer Science, a subfield called Human–Computer Interaction (HCI) frequently utilizes controlled experiments to assess the effectiveness of novel interaction techniques, algorithms, or systems. The focus of the thesis is specifically directed towards researchers in HCI and closely related disciplines, such as information visualization. HCI serves as an interdisciplinary field, situated at the intersection of numerous domains, including Psychology, Cognitive Science, Statistics, Design, and more. Consequently, methodologies and practices are often shared and adapted from these various fields, resulting in a diverse range of qualitative and quantitative methods. While the primary audience for this work consists of HCI researchers, it holds potential relevance for other scientific fields as well. This stems from the fact that the fundamental nature of experiments remains similar across different domains.

To start, I will provide an overview of the specific type of experiments that this work concentrates on. Following that, I will summarize the challenges associated with achieving good and accurate experimental design. Finally, I will conclude this section by providing an overview of the research objectives and hypotheses.

1.1 Designing the Right Experiment

HCI encompasses a multidisciplinary field where researchers employ diverse methodologies from various disciplines to gather empirical data and make novel contributions. Wobbrock and Kientz [2016] outlined seven common types of contributions within HCI, with particular emphasis on two types crucial for this work: empirical research

contributions and artifact contributions. Empirical research contributions, also known as empirical contributions, involve the gathering, analysis, and interpretation of qualitative or quantitative data to generate knowledge. On the other hand, artifact contributions involve the development of interactive systems, prototypes, or tools through design activities.

In many cases, artifact contributions are accompanied by empirical contributions to validate or compare them against existing practices. Providing empirical evidence is considered a valuable aspect of research, even though the validation or comparison may not be meaningful [Greenberg and Buxton, 2008]. For example, empirical contributions might focus on "the ease of system use" rather than "how useful a system is." Greenberg and Buxton [2008] discuss how the potential drawbacks of such usability evaluations and emphasize the importance of selecting an appropriate empirical method that aligns with the research question.

As a multidisciplinary field, researchers in HCI have a wide array of methodologies to choose from, including ethnographic observations, technology probes, interviews, diary studies, structured observations, quasi-experiments, and controlled experiments. Each method enables researchers to collect different types of data, which in turn support different types of claims. The selection of the appropriate method, in terms of its relevance to the research question, is equally as important as conducting the chosen method with high quality and rigor. Choosing the right method is a crucial prerequisite for designing the experiment correctly.

One of the methods employed in HCI research is the **controlled experiment**. In a controlled experiment, researchers manipulate specific factors to establish a causal relationship, while simultaneously controlling or minimizing the influence of other factors. By conducting the experiment in a controlled setting, researchers can reduce uncertainty regarding the interpretation of the results. The objective of a controlled experiment is to establish a causal relationship between a stimulus and its effect and to generalize these findings to the broader population from which the participants were recruited.[1]

[1] At the time of this thesis, Wendy E. Mackay, Joanna McGrenere, Chat Wacharamanotham, and I have been involved in a project focused on publishing structured observation as a practical methodology. As part of this project, the researcher assisted in surveying various empirical methods for comparison and specifically helped in

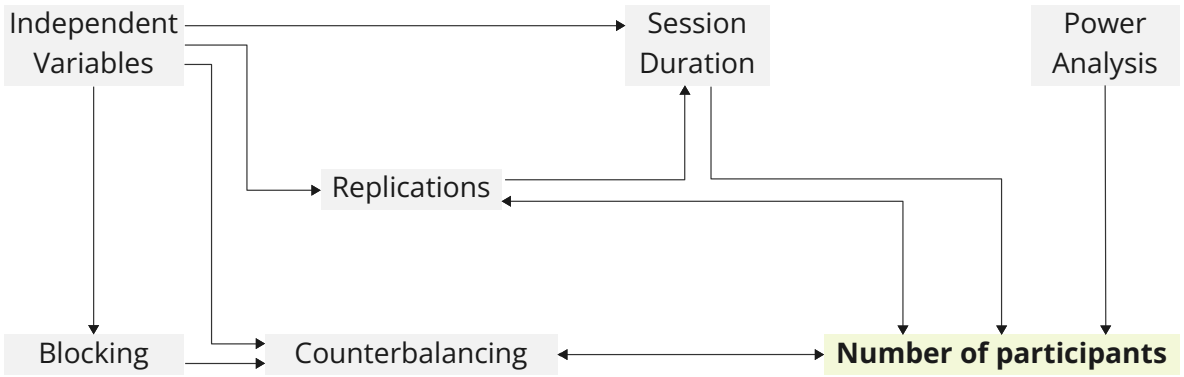


Figure 1.1: Relationships between different critical factors for experimental design. $A \rightarrow B$ indicates that a change in A might trigger a change in B.

1.2 Designing the Experiment Right

Designing a controlled experiment involves the specification of various parameters related to the experimental design. The resulting experimental design is then implemented to carry out the experiment. It is important to note that a well-designed experiment has the potential to provide strong empirical contributions, whereas a poorly-designed experiment can undermine the validity of the entire study. To begin, an overview of the pertinent design parameters and the process for designing an experiment will be presented. This will be followed by a discussion on the implications of good and bad experimental design. Finally, the focus will shift towards addressing the research questions.

1.2.1 Designing Experiments

In order to ensure a high level of quality and rigor in experiment design, researchers must carefully consider several design parameters, which involve making trade-off comparisons between different experimental designs. The first step is to operationalize the hypothesis by identifying one or multiple **independent variables**[2] (IV). Additionally, researchers need to determine the dependent variables, which represent the data collected during the experimental session. The next

comparing them with controlled experiments. However, it is important to note that this ongoing work is excluded from the scope of the present thesis.

[2] Also known as *factors* or *stimulus*.

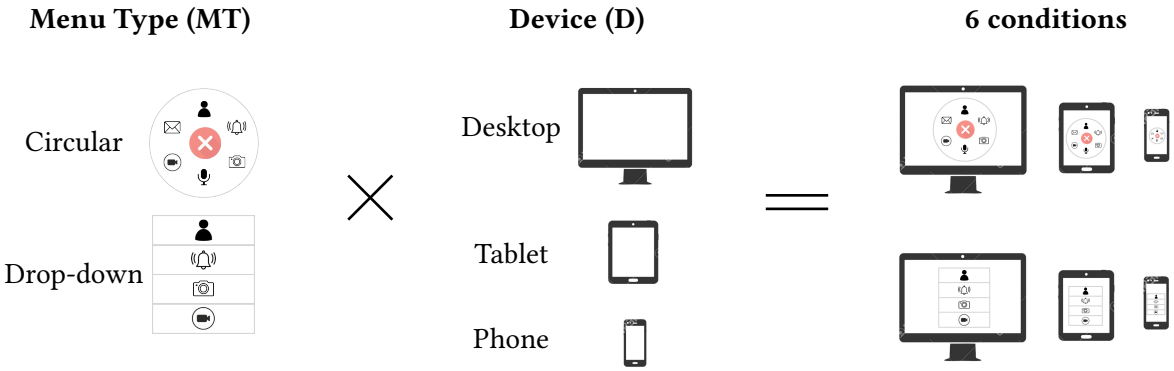
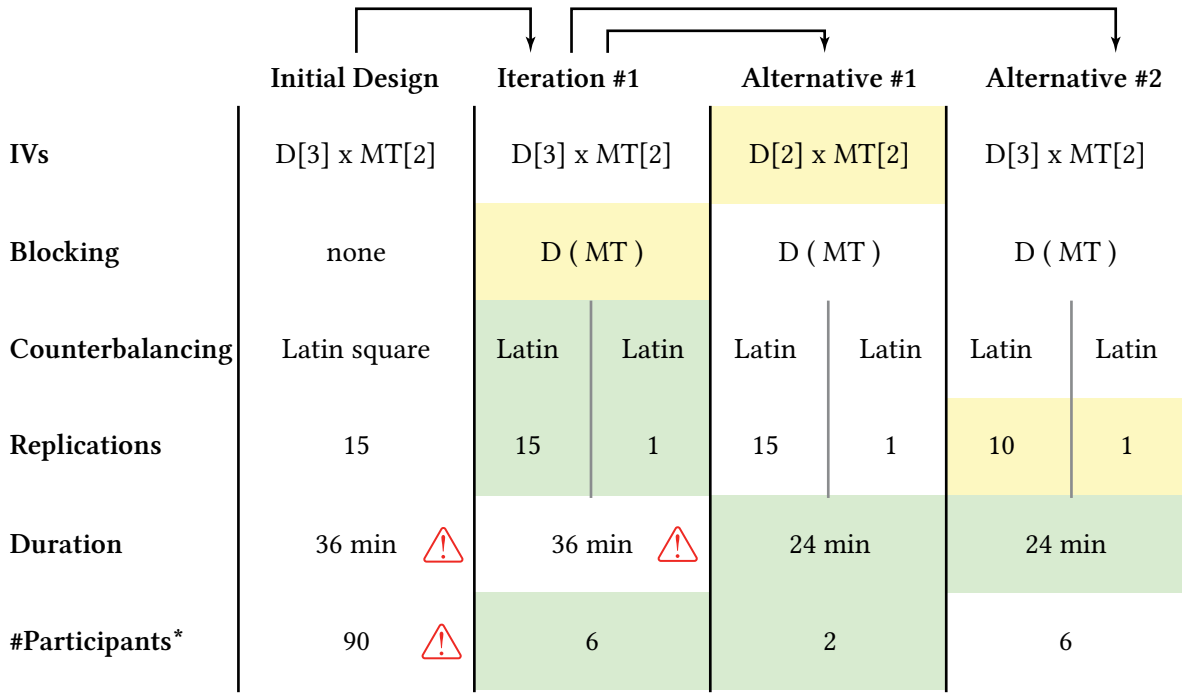


Figure 1.2: Two independent variables—Menu Type and Device—yield six conditions to compare with participants.

design considerations are **blocking** and **counterbalancing**. Blocking refers to how each combination of IVs is distributed among the participants, while counterbalancing involves systematically varying the order in which the different conditions or treatments are presented to minimize potential biases. In certain experiments, participants may be assigned multiple **replications** of the same combination of IVs. This repetition allows for aggregating measurements across replications to reduce noise in the data. Finally, researchers need to determine the **number of participants**—also referred to as **sample size**—to recruit for the experiment, which can be informed by a statistical **power analysis**.

Researchers often face challenges in balancing trade-offs between design alternatives due to the intricate relationships among experimental design parameters. Figure 1.1 illustrates the interdependencies between the aforementioned parameters. Modifying one parameter can have ripple effects on other parameters, leading to a complex web of relationships. These dependencies are non-trivial and often concealed, as predicting the impact of a parameter change is challenging due to the cyclic nature of these relationships. To better comprehend the complexity involved, let's consider an example of an experimental design.

Let's consider an example of experimental design to compare two Menu Types (MT): Circular and Drop-down, across three different Devices (D): Desktop, Tablet, and Phone (Figure 1.2). This setup gives us two **IVs**: Menu Type with two conditions (MT[2]) and Device with three conditions (D[3]). In total, there are $2 \times 3 = 6$ pairs of conditions that we want to compare. Each participant should receive all six con-



* indicates the multiple number of participants for a fully counterbalanced design.

Figure 1.3: Experimental design process from an initial design to two plausible alternatives. Yellow denotes the parameter that is changed by the researcher. Green denotes update parameters based on the researcher's change.

ditions to reduce the variation between participants. To lower noise in the data further, we set the **replications** to 15 so that each participant has to repeat each condition 15 times. To mitigate order effects, we will use Latin square counterbalancing. We estimate[3] the duration for each session per participant to be around 36 minutes. Latin square counterbalancing exposes each participant to the six conditions in a different order, so that each condition is seen only once and at a unique position. Using the Latin square counterbalancing and the parameters mentioned above, we calculate the required number of participants as $3 \times 2 \times 15 = 90$ participants.

There are two problems with the initial design (Figure 1.3) regarding the duration and the number of participants. One factor that constrains the number of participants is the budget researchers have available for the experiment. Let's assume the constraint is 30 participants. The initial design requires at least 90 participants, which ex-

[3]The following timings are used for the estimation: average duration per trial 20 seconds, delay after each trial 4 seconds, and delay after each block 1 second.

ceeds the available budget. We can address the problem of the number of participants as follows:

By blocking the conditions by Device, participants will complete all conditions for one device consecutively (Figure 1.3 Iteration #1). Therefore, a participant would complete all 30 Menu Type conditions (MT[2] $\times$ 15 replications) with the Desktop before moving on to the Tablet and finally the Phone. The design is now suitable for a multiple of 6 participants. With our budget, we could recruit either 24 or 30 participants.

The duration of 36 minutes is quite long for such an experiment. This might lead to reduced participant performance in later trials due to fatigue. Ideally, the duration should be shorter, preferably less than 30 minutes, with 25 minutes being even better to minimize fatigue effects. We can address the duration problem by considering either of the following two possibilities:

1. We can drop the Tablet condition from the IV Menu Type (Figure 1.3 Alternative #1). By removing one condition, the duration of the experiment would be reduced to approximately 24 minutes. However, it is important to note that this alternative would also eliminate one of the desired conditions from the experiment.
2. Another possibility is to reduce the number of replications to 10 (Figure 1.3 Alternative #2). This would also result in a duration of around 24 minutes.However, it is worth considering that reducing the number of replications increases the likelihood of noisier data, which can potentially compromise the statistical results.

Now, the researchers need to decide between the two alternatives by evaluating their respective trade-offs. Conducting a pilot study could provide additional insights and help determine the feasibility of the reduced number of replications. For example, the researchers could opt to run a small study comparing only the Tablet and Phone conditions to assess if there are any significant differences between them. Alternatively, they could collect data from a few participants during a pilot study to evaluate the level of noise present in the data.

This example provides a brief glimpse into the process and decision-making that researchers encounter when designing experiments. In

reality, experiments can be much more complex, involving multiple experimental conditions and stricter constraints on factors such as the number of participants, replications, and power analysis. Exploring the parameter space and carefully considering these design aspects is a time-consuming yet crucial task. Designing a good experiment is of utmost importance as it lays the foundation for obtaining reliable and valid results.

1.2.2 Consequences of Poorly Designed Experiments

A poorly designed experiment can undermine the entire study due to various factors such as reduced statistical power, challenges in replication, and the production of inaccurate or unreliable results.

Statistical power refers to the probability of detecting an existing effect in the population, given a specific sample size. Ignoring the concept of statistical power in experimental design can compromise the validity of the study, as it may lead to insufficient sample sizes or inadequate data collection that hinder the detection of effects. It is crucial to recognize that the absence of a statistically significant effect in the results does not necessarily mean the effect does not exist; it could be a result of low statistical power, leading to false negatives where true effects are missed. Therefore, ensuring an appropriate sample size and sound experimental design is important to minimize false negatives and increase the likelihood of detecting significant effects.

A crucial aspect of experimental design is the management of extraneous variables, which are factors that can potentially influence the relationship between the independent and dependent variables. Examples of such factors include incorrect counterbalancing, which can result in order effects that render the collected data unreliable, or inadequate blocking, which may lead to unintended order or learning effects. Failure to properly control for these extraneous variables can confound the collected data, introducing inaccuracies and reducing the reliability of the results.

Errors in experimental design can encompass various mistakes, such as incorrect counterbalancing, which can introduce order effects that render the collected data invalid. Additionally, incorrect blocking can result in unwanted order or learning effects that impact the reliability of the results. Moreover, an inadequate number of participants can

lead to an experiment with low statistical power, reducing the ability to detect meaningful effects.

A **replication study** aims to recreate an experimental setup as closely as possible, including aspects such as experimental design and participant recruitment. In some cases, slight variations may be introduced to examine if similar results can be obtained under slightly different conditions. Replication is a vital component of scientific research, but it can be hindered by poor experimental design. Difficulties in replication arise when the original study lacks thorough documentation, insufficiently describes the methodology, or fails to adequately control for external factors. These deficiencies in experimental design and transparency in reporting have contributed to a replication crisis in fields such as Psychology [Maxwell et al., 2015], Neuroscience [Marek et al., 2022], and Medicine [Ioannidis, 2005]. In many cases, researchers have struggled to reproduce the results of previously published studies due to insufficient information about the experimental design and analysis procedures [Goodman et al., 2016].

Ultimately, poor experimental design can have significant consequences for the credibility of both the research and the researchers involved. Flawed study design, incorrect data collection or analysis, and a lack of proper controls can all undermine the reliability and validity of the findings. When these issues arise, trust in the research and the researchers may be compromised. This loss of trust can hinder the researchers' ability to secure funding or have their work published in reputable venues. It is therefore crucial to exercise the utmost care and attention when planning and conducting an experiment to maintain the integrity of the research and the researcher's reputation.

1.2.3 Research Questions

In summary, designing a controlled experiment poses challenges due to the interdependencies between different parameters and the criticality of implementing correct and rigorous experimental design. Neglecting these considerations can lead to negative consequences for the research outcomes and the researchers involved. This leads to the first research question:

> **RESEARCH QUESTION 1:** How can researchers be supported when designing controlled experiments?

I investigate RQ1 in Chapter 2 by examining specific challenges that researchers encounter during the experimental design process. The goal is to develop a tool that facilitates the comparison of trade-offs between different design alternatives.

One of the most critical parameters in experimental design is the number of participants, also known as the sample size. Insufficient sample size can lead to inconclusive results and raise questions about the validity of the experiment. *A priori* power analysis allows researchers to compute a reasonable sample size based on an effect size from previous research, pilot data, or local standards within the field. However, the challenge lies in accurately estimating an effect size that would result in a sensible sample size, which leads to the second research question:

> **RESEARCH QUESTION 2:** How can researchers be supported when conducting *a priori* power analyses to inform the sample size?

I investigate RQ2 in Chapter 3 by developing an application that facilitates the exploration of contributing factors in power analysis and enables researchers to validate the insights they can gain from it.

Fisher [1937] and Neyman and Pearson [1928] established fundamental rules that form the basis of modern statistics. One of these rules states that the sample size should be predetermined before data collection and should not be analyzed until after the entire sample size is collected. However, this rule posed challenges in the context of medical trials, where large sample sizes, potential health risks for participants, and long durations made it impractical. As a result, researchers have developed methods that allow for ongoing statistical analysis while data is being collected [Armitage et al., 1969, Pocock, 1977, Bauer, 1989]. These methods have been adapted to various fields, including [Lakens, 2014b] and Neuroscience [Feder et al., 1991]. In HCI research, which often deals with novel technologies and interaction techniques, determining an appropriate sample size based on prior research remains a challenge. This leads to the third and final research question:

> **RESEARCH QUESTION 3:** How can researchers in HCI utilize a more flexible approach to plan sample sizes for controlled experiments?

I investigate RQ3 in Chapter 4 by providing the theoretical context for more flexible sample size decisions in HCI research. Additionally, I develop R templates and a web application that enable researchers to implement and conduct these flexible sample size calculations.

These three research questions and projects are connected by the theme of sample size planning with counterbalancing design. The first project, *Touchstone2*, focuses on the counterbalancing design and includes a simple prototype for determining a reasonable sample size. This prototype requires the user to specify a numeric value for the expected strength of the effect to be investigated, known as the effect size. However, we found that this specification is challenging for the user, even with statistical training.

The second project, *Argus*, addresses this issue by breaking down the effect size into smaller components. These smaller components are easier for the user to estimate based on prior work or knowledge about the expected data. However, in some cases, there may still be uncertainty in estimating these smaller components, which can result in a sample size that is either too large or too small.

The last project, SPEED, combines the learnings from the previous two projects and enables more flexibility in the sample size decisions. With SPEED, users can conduct statistical analyses during the data collection and have the option to stop the experiment early if certain criteria are met. All three projects follow a natural progression, and there is an opportunity to incorporate the lessons learned from SPEED back into *Touchstone2*.

These three projects have either been published or are currently undergoing revision for publication. Each chapter includes a publication statement that acknowledges my involvement and contribution to the respective project.

In conclusion, controlled experiments play a crucial role in providing empirical validation for artifact contributions in research. However, it is essential to ensure that conducting an experiment aligns with the research question at hand. When experiments are deemed appropriate, it is the responsibility of researchers to adhere to best practices in terms of rigor, transparency, and reproducibility to ensure the production of high-quality results. The objective of this work is to support HCI researchers in this process by identifying and addressing the

challenges they face and developing tailored artifacts that meet their specific needs.

Chapter 2

Touchstone2: An Interactive Environment for Exploring Trade-offs in HCI Experiment Design

Touchstone2 offers a direct-manipulation interface for generating and examining trade-offs in experiment designs. Based on interviews with experienced researchers, we developed an interactive environment for manipulating experiment design parameters, revealing patterns in trial tables, and estimating and comparing statistical power. We also developed TSL, a declarative language that precisely represents experiment designs. In two studies, experienced HCI researchers successfully used *Touchstone2* to evaluate design trade-offs and calculate how many participants are required for particular effect sizes. We discuss *Touchstone2*'s benefits and limitations, as well as directions for future research.

Publications: The work in this chapter is a collaboration with Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. The author is responsible for conducting and analyzing the interview study, the evaluations, and the design and implementation of *Touchstone2*. This work was published at CHI 2019 [Eiselmayer et al., 2019] and received a best paper award (top 1%).

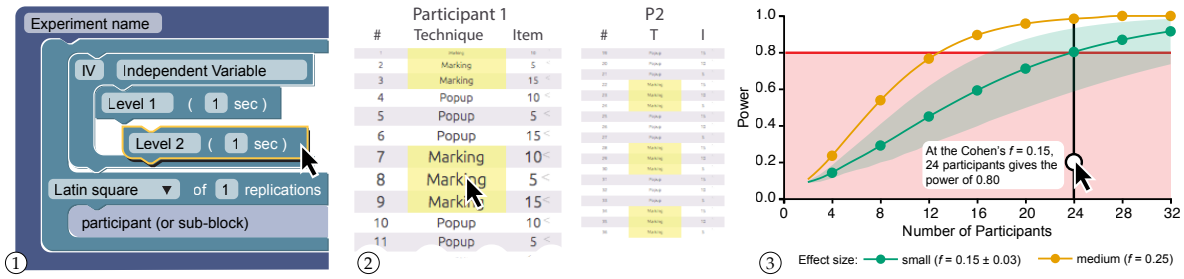


Figure 2.1: *Touchstone2* experiments consist of interactive "bricks" ① that specify independent variables, blocking, counterbalancing and timing, and generate an interactive trial table ② and an interactive statistical power chart ③.

2.1 Introduction

Human-Computer Interaction (HCI) researchers often compare the effectiveness of interaction techniques or other independent variables with respect to specified measures, e.g., speed and accuracy. Designing such experiments is deceptively tricky: researchers must not only control for extraneous nuisance variables, such as fatigue and learning effects, but also weigh the costs of adding more conditions or participants versus the benefits of higher statistical power.

Unfortunately, the problem is greater than simply helping individual researchers design experiments. The natural sciences face a "reproducibility crisis" — A recent survey of over 1500 scientists indicated that "more than 70% have tried and failed to reproduce another scientist's experiments." [Baker, 2016]. One explanation is the number of *researcher degrees of freedom*: the methodological decisions from study design up to publication [Simmons et al., 2011], including how many participants are recruited and assigned to which conditions [Wicherts et al., 2016]. Cockburn et al. [2018] argue persuasively in favor of pre-registering these decisions, in line with other scientific disciplines. However, to make this possible, the HCI community needs a common language for defining and sharing experiment designs. We also need tools for exploring design trade-offs, and capturing the final design for easy comparison with published designs.

Our goal is to help HCI researchers generate and weigh design choices to balance the inherent trade-offs among alternative designs. We present *Touchstone2* (Figure 2.1), a software tool for creating, comparing and sharing experiments that includes:

- *a visual environment* to manipulate experiment designs and their parameters;
- *a graphical interface* to weigh alternative designs and highlight trial table patterns;
- *an interactive visualization* to assess statistical power;
- *an online workspace* to compare and share designs; and
- *a declarative language*, TSL, to describe complex experiments with minimal constructs and operators.

After discussing related work, we present the results of an interview study that informed the design of *Touchstone2*. Next, we present the design rationale for *Touchstone2* and the TSL language, as well as the results of two workshops with HCI researchers to assess the interface. Finally, we discuss the benefits and limitations of *Touchstone2*, as well as directions for future research.

2.2 Related Work

This paper focuses on two aspects of experiment design: counterbalancing[1] and *a priori* power analysis. The research literature includes different conventions for representing experiment designs, and provides some software packages for ensuring counterbalancing and assessing power.

2.2.1 Representing Experiment Designs

Individual research disciplines use various techniques for optimizing experiment designs. For example, industrial manufacturing uses *Response surface design* [Box and Wilson, 1992] and the *Taguchi* method [Nair et al., 1992] for between-subjects designs. They treat product elements as experiment subjects and focus solely on determining the optimal number of levels for each independent variable. In the natural sciences, Soldatova and King [2006] created a computer-readable

[1] Statisticians use the more general term *randomization design*, which includes *counterbalancing*. The latter is more common in HCI. We use both terms interchangeably in this paper.

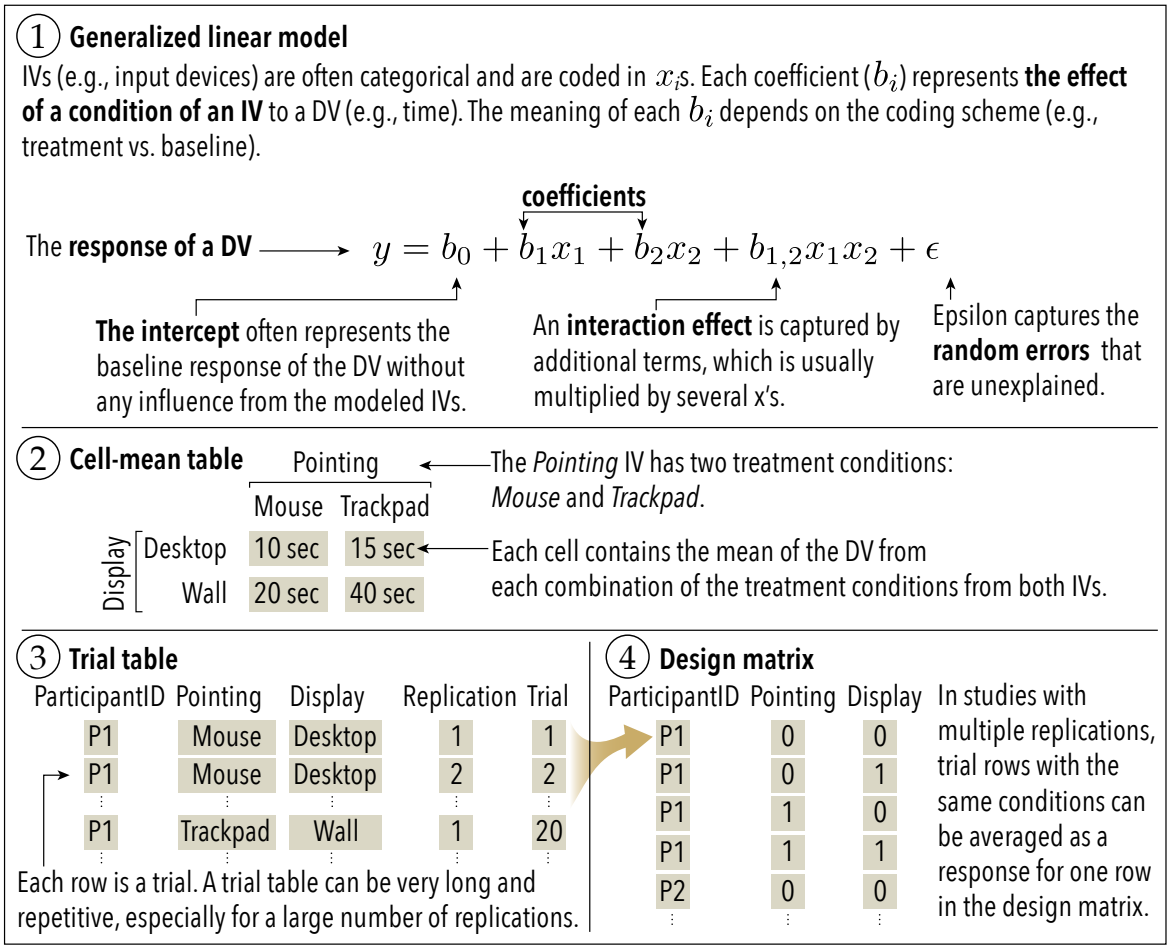


Figure 2.2: Four experiment designs representations [Cox and Reid, 2000][2].

ontology of scientific experiments (EXPO) that defines terms related to scientific discovery: *research, null* and *alternative hypotheses*, *independent* (IV) and *dependent variables* (DV), and *results*. This helped automate hypothesis generation and testing for yeast genomics experiments [King et al., 2009]. However, since experiments in this domain are restricted to simple Latin square designs, EXPO omits *blocking* and *counterbalancing*. Papadopoulos et al. [2016] present VEEVVIE, an ontology that describes Information Visualization data at the trial level, which unfortunately precludes specifying trial order.

[2]There are multiple ways to model the error term in a GLM. See dwoll.de/rexrepos/posts/anovaMixed.html based on [Wollschläger, 2017].

The statistical literature [Cox and Reid, 2000, Fisher, 1937] argues that experiment designs serve two primary goals: 1) explaining effects and 2) explaining the assignment of treatment conditions to subjects[3]. To explain effects, generalized linear models (GLM) determine the appropriate statistical procedures for data analysis (Figure 2.2 ①). Cell-mean tables ② summarize levels of dependent variables for each condition (often used in statistical reports and for power analyses).

Treatment condition assignments are often displayed as *trial tables*, with one trial per line ③, but their length and complexity make them cumbersome to manipulate. *Design matrices* provide two-dimensional representations of GLM coefficients, but without order information ④, as each row in a design matrix may correspond to multiple replicated trials. Text descriptions are also possible, but the lack of agreed-upon formats and minimum 'completeness' requirements increases the likelihood of incomplete or ambiguous experiment descriptions, especially within the page limitations required by publishers. We argue that comparative exploration of experiment designs requires a compact, yet flexible, formal specification of how treatment conditions are assigned to each participant.

2.2.2 Software for Specifying Counterbalancing

Counterbalancing a design is the process of assigning treatments to experiment units, e.g., participants. Experiments using a within-participant factor must counterbalance the treatment order to avoid systematic errors, minimize random errors, and ensure that interaction effects—if present—are captured [Cox and Reid, 2000]. Some statistical software packages, e.g., *JMP DOE* [SAS Institute Inc., 2016], *Design-Expert*[4], and the R package `skpr` [Morgan-Wall and Khoury, 2018] support part of this process. Experimenters must specify a GLM in order to generate trial tables with ordered sets of treatment conditions per participant. The IV levels are then optimized for maximum efficiency in large-scale, between-subjects experiments. However, most HCI experiments are small scale, with few participants [Kay et al., 2016b], and often include within-participant factors.

The `crossdes` R package [Sailer, 2013] generates trial tables and tabulates treatment frequencies by row, column, or concurrence, but only

[3] *Subjects* is the statistical term; we use *participants* for human subjects.
[4] jmp.com, statease.com

for within-subject designs. Each system offers a wizard-style dialogue for entering parameters. Some include examples, but few are directly relevant to traditional HCI experiments and none support comparing alternatives.

Both *Touchstone* [Mackay et al., 2007] and later *NexP* [Meng et al., 2017] were designed explicitly for HCI experiments that assess how human participants interact with specific technologies. Both offer novice researchers step-by-step instructions, with templates and menus to gather the parameters needed to generate a trial table. The *Touchstone design platform* leads users through a series of screens that specify independent variables and levels, blocking, counterbalancing, and timing. In-context help encourages users to evaluate potential negative consequences of particular decisions. The *Touchstone run platform* presents the resulting counterbalanced sets of trials to experiment participants. *NexP* offers an alternative question-answer approach to enter experiment design parameters. Both systems help users weigh the pros and cons of various decisions, but are designed for tweaking one design at a time, rather than systematically comparing alternatives. Neither offers a direct manipulation interface for generating experiment designs, nor an underlying declarative language for uniquely specifying each experiment.

2.2.3 Software for *a priori* Power Analysis

The HCI literature typically sets alpha levels to 0.05, lowering the risk of *false alarms*, i.e. Type I errors that claim an effect that does not exist. However, HCI experiments are often small, with only 12–16 participants. While these may detect large effect sizes, e.g., Bubble cursor's [Grossman and Balakrishnan, 2005] 30% speed increase, they significantly increase the probability of *misses*, i.e. Type II errors that do not find a real effect (Figure 2.3).

An *a priori* power analysis[5] lets experimenters determine the number of participants necessary to detect an effect of a specified size, given a significance criterion. Several calculators[6] and R packages, such as `pwr` [Champely et al., 2018], support power analysis. G*Power [Faul et al., 2007], currently the most comprehensive such, provides a form

[5] Shortened to *power analysis* in the paper

[6] For example http://www.macorr.com/sample-size-calculator.htm and http://www.dssresearch.com/KnowledgeCenter/toolkitcalculators.aspx

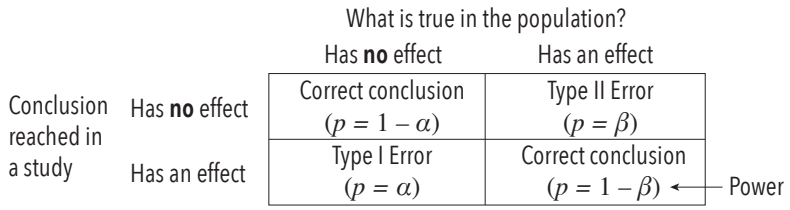


Figure 2.3: Type I and Type II errors, statistical power.

to enter the above parameters and calculates the minimum sample size. The resulting *power chart* shows relationships among sample size, power, and effect sizes, helping users assess the trade-offs between the benefits of additional power (detecting smaller effect sizes) and the cost of adding participants. No current HCI experiment design platform offers power analysis.

We argue that existing HCI experiment design platforms should be extended to support generating and visualizing alternative designs, based on randomization, power analysis, and other factors. This requires a common format for representing experiments, so they can be replicated and shared within the HCI community.

2.3 Interview Study

Prior to designing the *Touchstone2* interface, we investigated how experienced researchers currently design experiments: What challenges do they face and how do they resolve them?

2.3.1 Participants

We recruited 10 researchers who had designed, run and published one or more controlled experiments: 2 post-docs, 7 Ph.D. students and 1 graduate assistant, in Economics (1), Biology (1), Psychology (2) and HCI (6).

2.3.2 Procedure

We interviewed participants at work for 30-60 minutes, using the critical incident technique [Mackay, 2002]. We asked them to describe, step-by-step, the design of their current or most recent experiment, including any relevant tools or artifacts, e.g., spreadsheets. We probed for associated tasks, e.g., how they counterbalanced conditions across participants.

2.3.3 Data collection

We recorded audio (5) and hand-written notes (5). We took pictures of whiteboards and copied participants' hand-written notes, printed documents, scripts or spreadsheets used to create or communicate their designs.

2.3.4 Results

Participants highlighted the following design challenges:

Time constraints (8/10): P3 works with small children with short attention spans—so sessions can last at most five minutes. P9's pointing experiment was limited to 30 minutes to avoid fatigue.

Weighing design alternatives (6/10): P8 ran multiple pilot tests over four months that detected subtle, confounded learning effects. She ran a between-participants part to avoid learning effects and a within-participants part to let them compare the techniques. This required 27 participants, which was costly to recruit and run.

Counterbalancing problems (6/10): P4 spent several days unsuccessfully using a spreadsheet to generate a Latin square for a complex experiment. Despite the color-coding, his advisor was unable to verify his table and ended up recreating it from scratch, using her own counterbalancing method. P8 discovered a counterbalancing error at the third level of an independent variable *after* running her experiment. Fortunately, a post-hoc analysis showed no significant carryover effect. P9 created a trial table with a Python script but was not sure if it was counterbalanced correctly.

Representing experiment designs (7/10): P3 sketched her design on paper and on a tablet, with figures created in PowerPoint and Word, and P6 and P7 drew their designs on paper to get feedback. All had to recreate these representations after the design was changed.

Power analysis to select sample size (4/10): None of the HCI researchers used power analysis to choose the number of participants. Instead, they used the "at least 12" rule of thumb for small-n statistics, plus whatever was necessary for correct counterbalancing. Non-HCI participants treated power analysis as a suggestion and made adjustments later. For example, P1 added extra participants in case some dropped out of his online experiments. Others preferred smaller sample sizes due to restricted access, e.g., P2's studies of hospital employees; or the cost of samples, e.g., P10's studies of RNA sequences. P3 recruited as many children as possible and conducted post-hoc power analyses to demonstrate statistical power.

2.3.5 Discussion

We found that participants face numerous constraints, some predictable, e.g., P3's limited session time; some emergent, e.g., P8's discovery of a learning effect. They struggle to weigh the costs and benefits of different parameters and lack a standard way to represent and thus communicate their experiments. They also lack reliable methods for generating and verifying counterbalanced trial tables and assessing statistical power.

2.4 Designing *Touchstone2*

Touchstone introduced a streamlined process for counterbalancing trials [Mackay et al., 2007, Table 1], later adopted by *NexP* [Meng et al., 2017, Figure 1], with different views accessible in different tabs. The results from our interviews highlight the iterative and collaborative nature of the process, the multiplicity of artifacts generated to communicate designs (Figure 2.4), and the need to support power analysis (Figure 2.5).

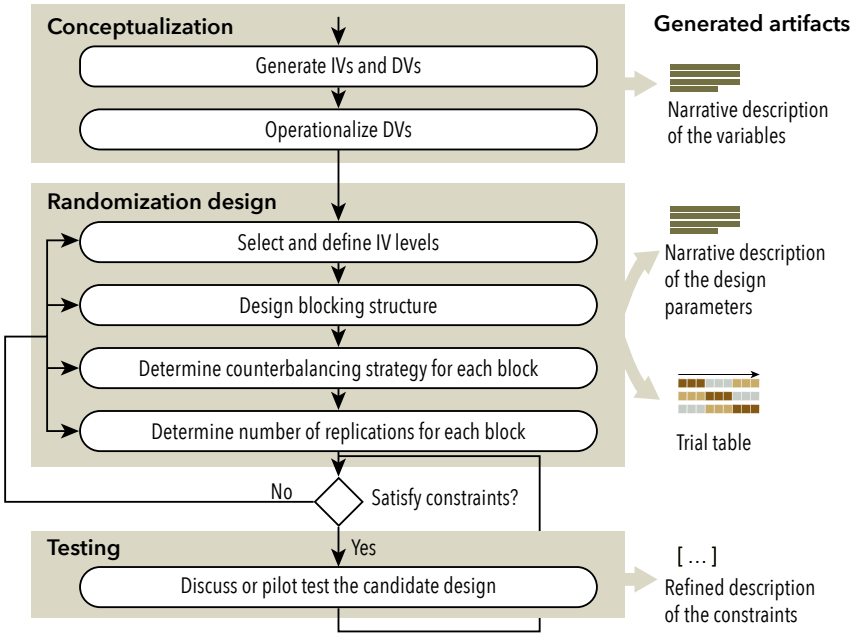


Figure 2.4: Counterbalancing is highly iterative: Multiple artifacts (right) capture, reveal, and communicate the design.

2.4.1 Counterbalancing Process

Researchers generate artifacts (Figure 2.4, right) to explore or communicate experiment designs, testing each candidate against constraints, e.g., number of participants or maximum session duration. Such constraints are often initially ill-defined, so researchers refine them based on pilot tests or suggestions from colleagues, in order to fully operationalize the design. Changes in earlier steps of the process affect later steps. For example, adding one level to an IV forces regeneration of the entire trial table. Both *Touchstone* and *NexP* let users repeat the operationalization step to automatically generate new trial tables. However, users must essentially start over if they make changes after importing a trial table into a spreadsheet to explore counterbalancing strategies or share with colleagues. *Touchstone2* therefore supports multiple parallel designs for easy comparison.

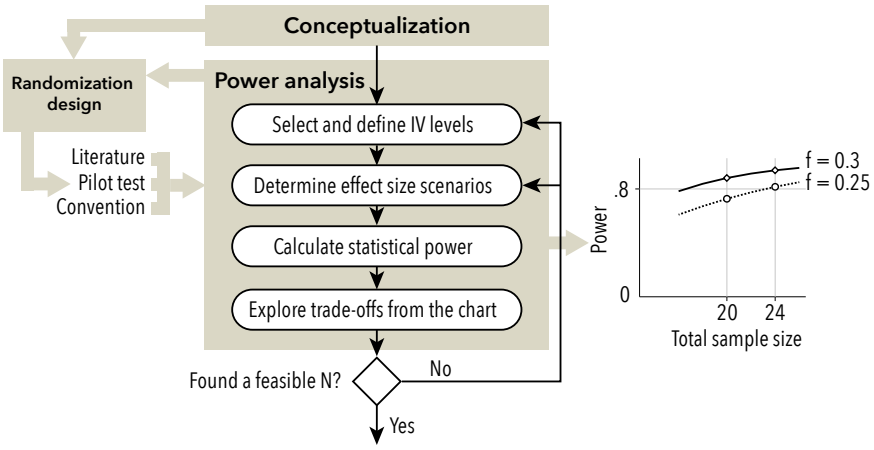


Figure 2.5: Power chart: Compare several possible effect sizes.

2.4.2 Power Analysis Process

Statistical power $(1 - \beta)$ is the probability of detecting a real population effect from the participants sampled in an experiment. This is computed from the sample size N[7], probability α of Type I errors[8], and *effect size*[9] in the real population. Studies with high statistical power are more likely to detect smaller effect sizes, but require larger numbers of participants.

Determining the experiment's sample size requires α and $1-\beta$ thresholds, usually .05 and .80 [Cohen, 1988, p. 56], and estimating the effect size (Figure 2.5). The latter is difficult and may discourage users from conducting a power analysis [Lipsey, 1990, p. 47]. Indeed, "*power analysis cannot be done without knowing the effect size in advance, but if we already know the size of the effect, why do we need to conduct the study?*" [Murphy et al., 2014, p. 17].

To cope with this conundrum, researchers usually visualize the relationships among N, power, and possible effect sizes in a *power chart* (Figure 2.5, right) to weigh the benefits of more power against the cost of more participants. In Figure 2.5 (left), increasing the sample size

[7] Number of participants
[8] Claiming an effect when one does not exist.
[9] How much DVs (measures) change according to different IV levels.

from 20 to 24 makes it easier to correctly detect a smaller effect size of 0.25 instead of 0.3.

Power analysis may be conducted either in parallel or after counterbalancing, depending on whether effect sizes are known, either from the literature or prior work. If such data is missing, researchers must either guess or run a pilot study. Not surprisingly, few HCI researchers run power analyses. Of 665 *CHI 2018* papers we examined, 519 include the term "experiment". Of these, 111 mention counterbalancing, but only five mention power analysis for choosing sample size. Our interviews indicate that, even though some HCI researchers know about power analysis, few use it, which increases the likelihood of missing small effects. *Touchstone2* facilitates power analysis, which helps researchers assess the risks of low power and make better-informed choices.

2.5 *Touchstone2*

The goal of *Touchstone2* is to facilitate exploration of experiment designs. We describe the user interface for specifying and comparing alternatives according to diverse criteria, e.g., randomization strategies (counterbalancing, blocking, replication), session length, and statistical power. Next, we describe the TSL language for specifying experiment designs.

2.5.1 *Touchstone2* User Interface

Each experiment consists of nested *bricks* that represent the overall design, blocking levels, independent variables, and their levels. Experiments can be assembled from scratch or cloned from a template, e.g., a [2x3] design. Parameters such as variable names, counterbalancing strategy and trial duration are specified in the bricks and used to compute the minimum number of participants for a balanced design, account for learning effects, and estimate session length. An experiment summary appears below each brick assembly, documenting the design.

In Figure 2.6, Design ① is a [2x3] within-participants design to compare menus, where TECHNIQUE has two values: POPUP and MARK-

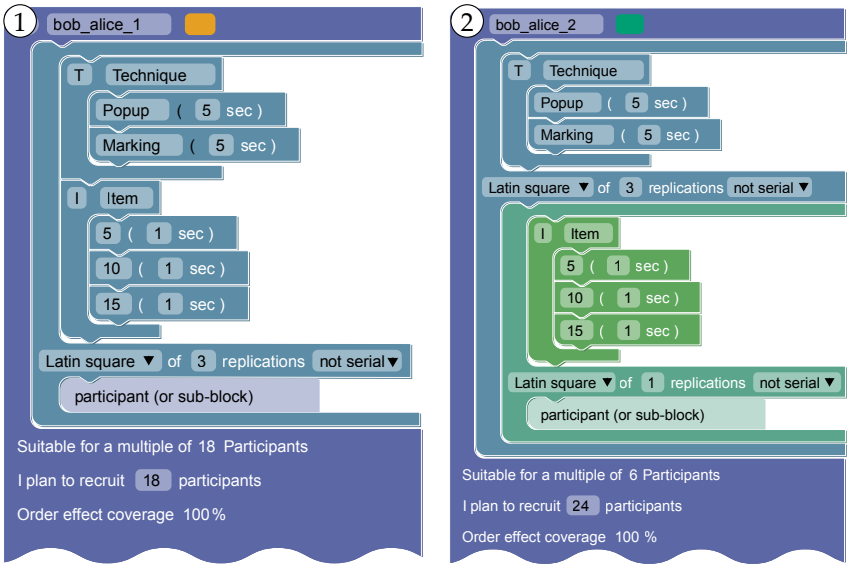


Figure 2.6: Two blocking strategies for a [2x3] within-participants design to compare POPUP and MARKING menus.

ING, and ITEM has three values: 5, 10, and 15. Trials are replicated three times. Design ② is blocked by technique, using a Latin square.

Counterbalancing

Users arrange bricks in a 2D workspace to enable side-by-side comparisons of alternatives. For example, in Figure 2.6, Design ① features a Latin Square brick that contains two bricks, one for each IV. This counterbalances all variables within the same blocking level, resulting in a balanced design for multiples of 18 participants. Design ② uses two Latin Square bricks. The brick that contains the *Item* IV is nested inside the brick that contains the *Technique* IV. This creates a blocked design, where trials are grouped by Technique level (Figure 2.7). As a result, the design is now balanced for multiples of only six participants.

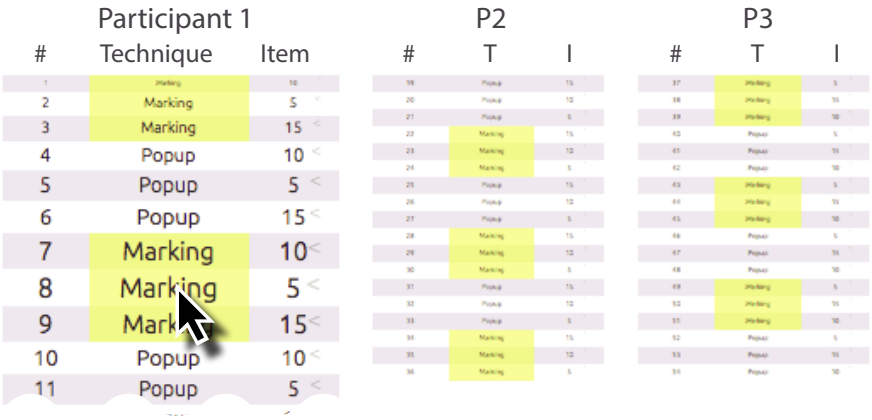


Figure 2.7: Trial Table Inspection with Fish-eye View

Inspecting Trial Tables

Manipulating bricks immediately generates a corresponding *trial table* (Figure 2.7) that shows the distribution of experiment conditions across participants. Trial tables are faceted by participant. The width and height of each table correspond to the numbers of participants and trials, respectively, to facilitate comparison.

Touchstone2 provides two tools for in-depth trial table inspection:

1. *Brushing* [Tweedie et al., 1996]: clicking on one or more cells highlights those corresponding to the same condition; clicking on one or more rows highlights those corresponding to the same combination of conditions.
2. *Fish-Eye Views* to show a TABLE LENS [Rao and Card, 1994] visualization: The trial table shrinks to an overview, magnified around the cursor for readability.

Users can easily compare among participants and among designs on one screen, and examine their trade-offs. For example, more independent variables will increase the study duration for each participant, hence the height of the table will be larger. Used together, these tools make it easy to inspect patterns of trial conditions and compare experiment designs. For example, Figure 2.7 highlights each MARKING level to show how they are grouped in consecutive trials.

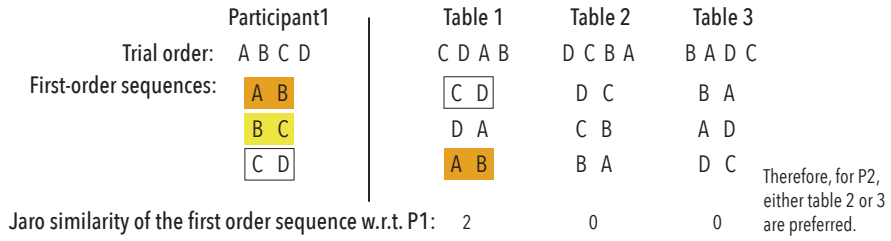


Figure 2.8: The Jaro similarity measure ensures maximum counterbalancing coverage for each successive participant.

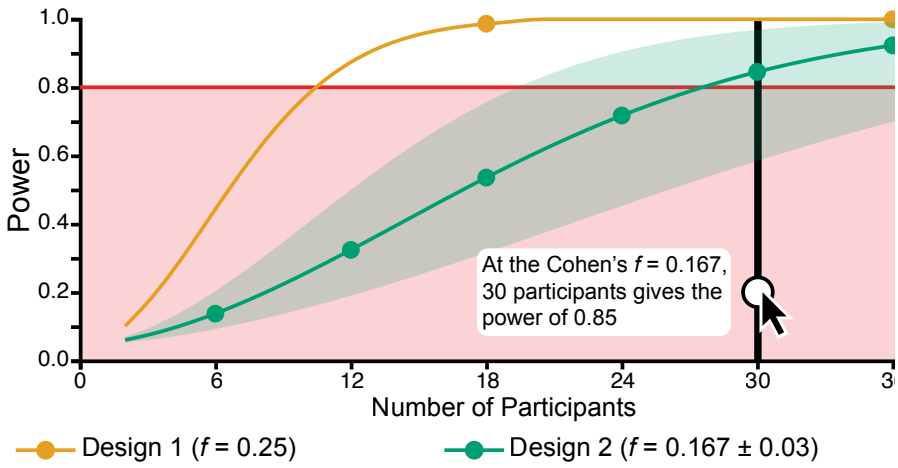


Figure 2.9: Power analysis: With 18 participants Design 1 is likely to find the effect. Design 2 needs 30 participants.

Touchstone2 orders trial tables so as to maximize counterbalancing coverage for each successive participant, in case too few participants are recruited or one drops out. Figure 2.8 illustrates this algorithm: Suppose we pick a trial table P_i for the i-th participant. The table for the next participant, P_{i+1}, is selected from those whose sequence of first-order effects are least similar according to the Jaro similarity measure (number of row-transpositions) [Jaro, 1989].

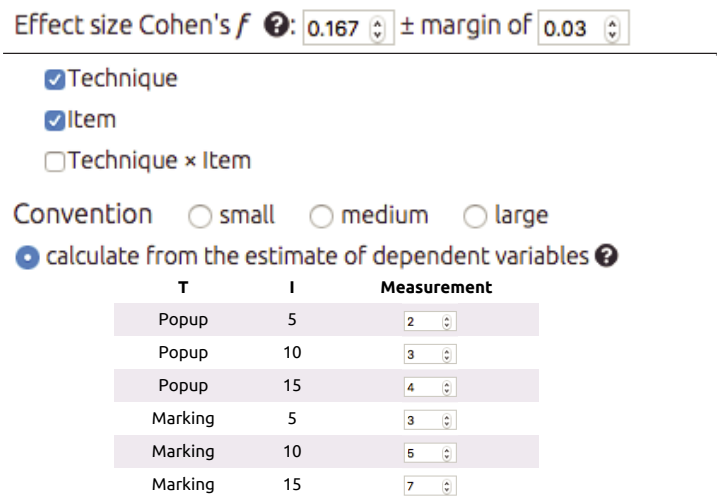


Figure 2.10: Calculating effect size from pilot data.

Power Analysis

Touchstone2 starts with a set of default parameters[10] and plots a power chart for each active experiment design in the workspace (Figure 2.9). Each power curve is a function of the number of participants, and thus increases monotonically. Dots on the curves denote numbers of participants for a balanced design. The pink area corresponds to a power less than the 0.8 criterion: the first dot above it indicates the minimum number of participants.

To refine this estimate, users can choose among Cohen's three conventional effect sizes [Cohen, 1988, Chapter 8], directly enter a numerical effect size, or use a calculator to enter mean values[11] for each treatment of the dependent variable (often from a pilot study). Users can select the factors and interactions to include in the power calculation, which automatically adjusts the degrees of freedom used to determine power. By default, all factors are included without interactions (Figure 2.10).

[10] Cohen's medium effect size $f = 0.25$, Type II error $\beta = 0.2$, Nonsphericity correction $\epsilon = 1$. These default parameters can also be globally customized.

[11] For skewed data, e.g., task completion time, users can instead input a more robust central tendency estimate, e.g., geometric mean or median. We leave non-interval data, e.g., Likert items, for future work.

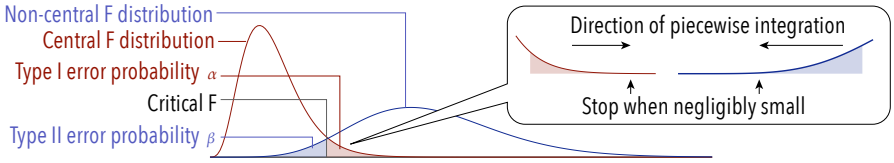


Figure 2.11: In the power calculation, the direction of integral calculation were optimized for responsiveness.

The power chart is a common representation in power analysis which is also available in G*Power. In *Touchstone2*'s chart, the user can compare *multiple* experiment designs and *interact* with them: Hovering the mouse cursor displays a vertical ruler that snaps to valid sample sizes. Users can click on any experiment in the workspace to highlight the associated curve. Users can also specify a margin of uncertainty around the estimated effect size. The power chart then displays an error band showing the corresponding margin of error on the power calculation.

Touchstone2 uses Cohen's f as the measure of effect size as it applies to multiple types of experiments, including within-participant and mixed designs[12]. Type I and Type II error rates (α, β) are calculated by integrating the probability distribution of a central and a non-central F distribution (Figure 2.11). Since this calculation[13] can reduce responsiveness, we optimize the numerical integration by adjusting the direction of each iteration according to the overlap between the distributions (Figure 2.11, callout). On average, each curve can be calculated in 300 ms with a single thread running on a 2.5 GHz Intel Core i7 processor. We also spawn one thread per curve to parallelize the calculation.

[12] According to the experiment design and selected effects (Figure 2.10, top), *Touchstone2* adjusts how the means values (Figure 2.10, bottom) are aggregated and how the degrees of freedom in the F distributions are calculated from the number of participants. See [Faul et al., 2007, Table 3] for detailed mathematical formulae.

[13] To produce smooth curves, we calculate power for sample sizes between 1 and 50. At each step, we integrate the probability distribution piecewise, in 0.1 increments, and adaptively increase precision 10 times until the resulting curve increases monotonically.

Online Help

Touchstone2 displays contextual help to the right of the screen, encouraging users to weigh specific trade-offs relevant to their current design. Note that *Touchstone2* is not intended as a standalone tutorial or replacement for an introductory course and assumes a basic understanding of experiment design. Of course, *Touchstone2* can complement an HCI experiment design course.

Collaboration and Sharing

Workspaces can be shared asynchronously using a simple web server. Users can export their trial tables in CSV format for use with statistical or other software, e.g., to log data. Users can publish experiments using the TSL format (described below), which contains a concise description of variables and nesting. Users can also export an entire workspace, including spatial placement of the bricks, comments, and power analysis input parameters, into an XML file. *Touchstone2* can export *Touchstone*-compatible XML files and load them into its run platform to present the experiment [Mackay et al., 2007].

Supported Platforms

Touchstone2 is implemented as a web application that works on SAFARI, CHROME, and FIREFOX. The code relevant to experiment design is written in 3477 SLOC of JavaScript with extensive use of Google's BLOCKLY library[14]. We debounce the change events within 200 ms before recomputing the trial table in a Web Worker[15] to avoid blocking the user interface. *Touchstone2* can be used locally or in conjunction with a lightweight web server (18 SLOC PHP script) for sharing designs.

[14] https://developers.google.com/blockly/
[15] https://www.w3schools.com/html/html5_webworkers.asp

2.5.2 Touchstone language (TSL)

The counterbalancing strategy specified by *Touchstone2* bricks is converted into a text specification using the Touchstone language (TSL), a domain-specific declarative language for describing randomization designs, e.g., counterbalancing. The TSL design goals are to:

1. Provide a concise and unambiguous description of randomization designs;
2. Cover a broad class of randomization designs;
3. Minimize operators for composing such designs; and
4. Reuse existing conventions as much as possible.

Each TSL experiment design is described by an assembly of experiment design blocks that specify the counterbalancing strategy, the independent variables and their levels, and the number of replications. For example, a Latin-square block with a 3-level IV DEVICE and four replications is written as:

```
<Latin(Device={M,T,J}, 4)>
```

Blocks can be assembled into a complex experiment design using four operators: `nest (A(B))`, `cross (A x B)`, `concatenate (< A, B >)` and `replicate (10 * A)`. For example, consider a mixed-design experiment with one between-subject factor[16]: POINTER (ACCELERATED, STATIC), and a within-subjects factor: DEVICE (MOUSE, TRACKPAD, JOYSTICK). This experiment tests different indices of difficulty `ID` with one training session and ten test sessions. In the training session, the order of the device is randomized, and the ID is fixed between 2 to 3. In the test session, both factors are counterbalanced with a Latin square. This experiment can be described in TSL as:

```
< Training = Between(Pointer = {A,S}, 1,
                     Random(Device = {M,T,J,R}, 2,
```

[16]Independent variables or IVs are also referred to as *factors*.

```
                                Fix(ID = {2,3}, 1))),
    10 * Between(Pointer = {A,S}, 1,
                    Latin(Device = {M,T,J,R}, 3,
                          Latin(ID = {2,3,5,6}, 1))) >
```

TSL can express within-subjects, between-subjects, and mixed designs. It implements four counterbalancing algorithms frequently used in HCI studies: Latin-square, complete permutation, random assignment, and fixed order. More sophisticated counterbalancing algorithms can be added as plug-ins. TSL also supports replications and multi-session designs, which are currently beyond the scope of the *Touchstone2* block-based interface.

The TSL generator is written in TypeScript[17] and compiled into JavaScript. The full TSL grammar comprises 12 production rules written in jison[18]. The generator can be used from the command line (as a Node.js application) or in a web application (as a JavaScript package) to generate a trial table from a TSL specification.

TSL offers a compact and unambiguous format for communicating experiment designs, and could be used to pre-register HCI experiments Cockburn et al. [2018]. The textual format allows changes to be easily identified with a *diff* tool and tracked with a version control system. The *Touchstone2* interface is more convenient for exploring experiment designs, and can both read and export TSL specifications.

2.6 Evaluation

We ran two evaluation studies. A workshop assessed the *Touchstone2* interface to see how well pairs of experienced researchers could counterbalance an experiment created by one partner and explore design alternatives. A second observational study focused on how individual participants assessed the statistical power of their earlier designs.

[17] https://typescriptlang.org
[18] https://zaa.ch/jison/

2.6.1 Workshop: Reproducing an Experiment

Participants

We recruited 17 experienced HCI researchers: 11 Ph.D. students, two post-docs and four faculty members.

Apparatus

Each team worked with an early version of *Touchstone2* on one of their personal laptops. This version supported within-participant designs, contextual help and fish-eye views of trial tables.

Procedure

16 participants worked in pairs, with at least one highly experienced researcher in each team. The remaining participant, a senior faculty member, worked alone. The workshop was conducted around a U-shaped table to let teams easily participate in the group discussion.

The workshop lasted approximately 90 minutes, beginning with a 15-minute introduction to *Touchstone2* and a description of the following tasks:

1. Choose your own current or recently published experiment;
2. Reproduce it with *Touchstone2*; and
3. Explore at least two variations of the experiment.

Participants had 60 minutes to work. Two authors observed the teams, answered questions about *Touchstone2* and noted any bugs, problems, desired features or suggestions for improvement. We encouraged participants to write any feedback or observations in the text area provided. Participants shared their impressions of *Touchstone2* in a final plenary discussion (15 minutes).

Data Collection

We collected logs of each team's experiment creation process, their final experiment design(s) and their written feedback, as well as the observers' notes.

2.6.2 Results

Most teams (8/9) successfully reproduced their chosen experiment in *Touchstone2*. (The unsuccessful team produced a simpler variation of their experiment instead.) The experiment designs that participants reproduced were relatively complex: Six teams reproduced experiment designs that involve three variables. Among these, half organized variables into two nesting levels, and the rest used three nesting levels. One team produced a design for four independent variables in two blocks. All teams used a Latin square counterbalancing strategy at least once. Two teams created a dummy independent variable to denote training vs. testing trials.

All teams adjusted parameters within each design, e.g., number of participants or counterbalancing strategies, and inspected how trial tables change. Most teams (6/9) created multiple versions of an experiment design ($Mdn = 2, Max = 4$). Two teams saved designs with different time estimates and numbers of replications. Two others produced versions with different nesting structures; one even split an independent variable into two variables at the same nesting level.

In seven teams, only one partner knew the experiment details. They mentioned that the visual representation of the experiment made it much easier to explain the design. They also mentioned that automatically updating trial tables encouraged them to explore more alternatives.

Two teams found it difficult to keep track of the reasons why they adjusted their design and suggested adding an annotation feature to document the process. Although some were interested in highlighting trial tables, teams that explored more complex designs emphasized the need for highlighting the pattern of *all* conditions in a row. We added these features to *Touchstone2*.

2.6.3 Observational Study: Analyzing Power

Participants

Ten individuals from the workshop were available for the second study: 5 Ph.D. students, 2 post-docs (P2, P10) and 3 faculty members (P6–8).

Apparatus

Participants worked on a computer with a revised version of *Touchstone2* that included power analysis. We uploaded the participant's final experiment design from the workshop.

Procedure

Sessions lasted approximately 30 minutes. The experimenter presented the interface changes in *Touchstone2* (v0.2), using one of the participant's experiment designs as an example, and explained the concept of statistical power, when necessary. Participants were then shown how to toggle the power analysis mode.

Participants were asked to replicate their experiment, first reassessing the current design and then determining the appropriate number of participants. We used a think-aloud protocol, with periodic reminders. At the end of the session, the experimenter conducted a semi-structured interview. Questions included how statistical power analysis affected the number of participants they decided to recruit, as well as comments about the user interface.

Data collection

We screen recorded 9/10 sessions and audio recorded all 10 interviews. The interviewer and an additional silent observer also took field notes.

2.6.4 Results

We selectively transcribed the audio and video based on field notes. Two authors analyzed the transcripts using thematic analysis [Braun and Clarke, 2006] using a bottom-up approach, i.e. without predefined research questions.

Attitude

P1–4 were explicitly skeptical of power analysis because of (1) the difficulty in recruiting participants (P1–3), (2) the existence of minimum sample size conventions (P3,P4): *"in my statistics courses, the rule is if you want to say anything that is relevant [sic] grab 30 or more."* (P4), and (3) the lack of incentive to run power analyses (P2,P4): *"until it is mandatory in a submission I would never do it"* (P2)). However, P2–4 mentioned its benefits while using *Touchstone2*.

Interpreting power charts

Five participants actively interpreted the power chart. Three wanted the power *"above [the threshold of 0.8] because it's red"* (P2). Three noted the diminishing returns as the power curve starts to plateau: *"The curve also gives you information how worth it is to keep adding participants beyond [the plateau]"* (P5). Three said that power differences would influence their recruitment decisions: *"If recruiting participants is not very hard I would probably perhaps [add more]. It seems more sound."* (P10). One said she would use the power chart to justify recruiting fewer participants. *"If I am struggling [recruiting], I think the chart is useful to say OK, no."* (P3)

Four participants said that power analysis would help make *"a stronger case"* (P4) in their paper submissions, especially with small numbers of participants. As a reviewer, P4 would judge a paper with power analysis more favorably, although P6 was neutral about it.

Barriers to power analysis

Understanding standardized effect size was a barrier for 9/10 participants (one of them is even an expert in statistics). Five said that they do not know how to interpret standardized effect size: *"What would be the range of values that would normally be?"* (P2); *"What's the intuition behind that? [...] and it is related to a specific domain although for me it doesn't say much"* (P8, an expert in statistics) Of these, three are knowledgeable about simple effect sizes, e.g., percentage difference. Participants felt it would be cumbersome to manually fill in the cells in the cell-mean table (3/10), and asked about how to deal with outliers in the data (3/10). The two experts in statistics wanted greater transparency in how effect size is calculated.

2.6.5 Summary

These results suggest that *Touchstone2* encourages users to explore alternative counterbalancing designs. However, 5/9 teams iterated their designs within a single experiment brick assembly and did not take advantage of the ability to manage multiple designs in the workspace. A possible reason is that the trial table is updated immediately after a change, making it easy to spot the effect of the change. However, this loses track of earlier designs. We could address this by improving the interface for accessing historical versions, and by making it even easier to duplicate a design.

Although participants quickly understood the benefits of the interactive power chart, the costs of estimating and interpreting standard effect size proved to be a major barrier. We thus revised the *Touchstone2* interface to first present the power chart, using Cohen's medium effect convention, and then provided options for controlling effect size in increasing order of complexity (see Section 2.5.1). We also added an explanation about standardized effect sizes and their calculation in the context-sensitive help.

2.7 Discussion

Touchstone2 opens several directions for future research for both practical and statistical aspects of experiment design.

2.7.1 Default Parameters and Status Quo Bias

To calculate power, *Touchstone2* uses default parameters and Cohen's conventions [Cohen, 1988, Chapter 8]. These defaults allow us to clearly signify the presence and the importance of statistical power without first requiring additional input. Although these parameters are customizable in the *Touchstone2* user interface, users may leave them unchanged because of *status quo bias* [Kahneman et al., 1991]. We recognize the risk that *Touchstone2* might encourage blind adoption of certain conventions without reflection, just as with the .05 threshold for p-values in the NHST paradigm. However, we argue that this issue arises in the teaching of statistics and experiment design, as well as the peer-review process itself. We hope that *Touchstone2* can contribute to the conversation about these issues. Ultimately, the trade-off between supporting discoverability and the risk of oversimplification is beyond the scope of this work.

2.7.2 Statistical Significance and Power Analysis

Power analysis in *Touchstone2* is a practice under the null-hypothesis significance testing (NHST) paradigm. The theory of power analysis—regardless of the software tools—can be abused for *p*-hacking. Researchers may calculate power mid-experiment and add more participants until achieving statistically significant results. Despite this problem and other criticisms, conducting transparent and valid research under the NHST paradigm is still possible through preregistrations [Cockburn et al., 2018], transparent communication of the results [Dragicevic, 2016, Transparent Statistics in Human–Computer Interaction Working Group, 2019], and reporting effect sizes [Transparent Statistics in Human–Computer Interaction Working Group, 2019, Chapter 2]. *Touchstone2* also facilitates better NHST practices. For example, *Touchstone2* presents the relationship between the number of participants and statistical power prominently in the UI. It also facilitates calculating effect size from the results of pilot

studies or using effect sizes from the literature. (The HCI community has created several guidelines and discussion such as [Yatani, 2016, Transparent Statistics in Human–Computer Interaction Working Group, 2019].) We believe that these aids will persuade researchers to plan experiments with high statistical power instead of *p*-hacking.

2.7.3 Integrating Data Analysis

Experiment design is inextricably linked to data analysis: A plan to aggregate data influences the experiment design. For example, Fitts's law experiments may be susceptible to high variance between trials due to motoric noise. If multiple trial replications, i.e. the same user performing the same technique multiple times, are averaged before statistical analysis, the number of trials (from the counterbalancing design) will differ from the sample size (in the power analysis). Therefore, the researcher should consider a trade-off between adding participants vs. increasing the number of trial replications for each participant.

This highlights the need for a clearer link between experiment design and data analysis. We believe that TSL and *Touchstone2* offer a basis for integrating both processes.

2.8 Conclusion

Our primary goal is to improve the quality and reproducibility of HCI experiments by offering researchers a tool for specifying and comparing alternative experiment designs. High-quality experiments require trade-offs: For example, shorter experiments with fewer conditions are easier to analyze and more comfortable for participants but provide potentially fewer results. These trade-offs are particularly challenging for HCI researchers, who commonly use small numbers of participants and low-power statistical tests. Also, experiments are more likely to be reproducible when researchers have complete and unambiguous specifications of experiment designs, which may be unavailable in research papers due to the lack of common language and page limits.

In this paper, we present four contributions. First, an **interview study** reveals that experiment design is iterative and collaborative. Researchers create, revise, and exchange design specifications and trial tables. However, keeping them in-sync is tedious and error-prone. Researchers also weigh the cost of participants against the benefit of statistical power. Additionally, the cost of *calculating* statistical power itself is also weighed against the practicality of its outcome. In summary, researchers navigate the trade-offs not only about the design itself but also about their design process.

Based on these findings, we present ***Touchstone2***, a direct manipulation interface for generating, comparing, and sharing experiment designs. *Touchstone2* lets researchers assess experiment designs with four metrics: (1) learning effects, (2) session duration, (3) number of participants, and (4) statistical power. These metrics are supported by instantaneous feedback on trial tables and power charts as well as an interactive visualization for inspecting them. All are provided in an online sharable workspace.

To improve the reproducibility of experiments, we contribute **TSL, a declarative language for experiment designs** that can express a large class of designs with few constructs and operators. TSL lets researchers share their designs in a concise and unambiguous format. A design expressed in TSL can be imported into *Touchstone2*, and can generate a trial table with a command line. Other GUIs for experiment design can also use TSL as a backend. TSL could be integrated into future preregistration, review, and publication processes to reduce ambiguity of experiment designs. Future work may extend TSL to, e.g.,, provide natural language descriptions or alternative visualizations.

Touchstone2 was **evaluated in two studies**. Our results show that *Touchstone2* encourages experienced researchers to explore alternative experiment designs and to weigh the cost of additional participants against the benefit of detecting smaller effects.

Both *Touchstone2* and TSL are available as open source projects[19]. We hope that they will provide a foundation for creating a repository of HCI experiments that will act as a resource for researchers, students, and educators to learn from existing experiment designs, weigh the pros and cons of specific experiments, and ultimately contribute to the reproducibility of HCI experiments in the research literature.

[18] https://github.com/ZPAC-UZH/Touchstone2
https://github.com/ZPAC-UZH/TSL

Chapter 3

Argus: Interactive *a priori* Power Analysis

A key challenge HCI researchers face when designing a controlled experiment is choosing the appropriate number of participants, or sample size. *A priori* power analysis examines the relationships among multiple parameters, including the complexity associated with human participants, e.g., order and fatigue effects, to calculate the statistical power of a given experiment design. We created *Argus*, a tool that supports interactive exploration of statistical power: Researchers specify experiment design scenarios with varying confounds and effect sizes. *Argus* then simulates data and visualizes statistical power across these scenarios, which lets researchers interactively weigh various trade-offs and make informed decisions about sample size. We describe the design and implementation of *Argus*, a usage scenario designing a visualization experiment, and a think-aloud study.

3.1 Introduction

Determining *sample size* is a major challenge when designing experiments with human participants, e.g., in Information Visualiza-

Publications: The work in this chapter is a collaboration with Xiaoyi Wang, Wendy E. Mackay, Kasper Hornbæk, and Chat Wacharamanotham. The author shared responsibility for both think-aloud studies and the implementation. This work was published at VAST 2021 [Wang et al., 2021].

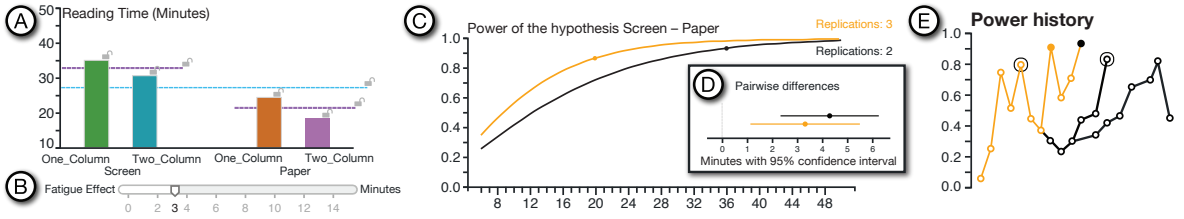


Figure 3.1: Argus interface: (A) Expected-averages view helps users estimate the means of the dependent variables through interactive chart. (B) Confound sliders incorporate potential confounds, e.g., fatigue or practice effects. (C) Power trade-off view simulates data to calculate statistical power; and (D) Pairwise-difference view displays confidence intervals for mean differences, animated as a *dance of intervals*. (E) History view displays an interactive power history tree so users can quickly compare statistical power with previously explored configurations.

tion (VIS) and Human-Computer Interaction (HCI) [Eiselmayer et al., 2019, Hornbæk, 2013, Lazar et al., 2017a]. Researchers want to save time and resources by choosing the minimum number of participants that let them reliably detect an effect that truly exists in the population. However, if they underestimate the sample size, i.e. the experiment lacks statistical power, they risk missing the effect – a Type II error. Researchers are also less likely to publish these negative or null results, the so-called "file drawer problem" [Rosenthal, 1979]. Researchers cannot simply add participants until the results are significant, which is considered a malpractice, and are strongly encouraged to preregister the sample size to increase the credibility of the investigation [Cockburn et al., 2018].

The sample size can be determined statistically with an *a priori* power analysis. However, this requires approximating the *effect size*, which quantifies the strength and consistency of the influences of the experimental conditions on the measure of interest. Estimating an effect size must account for the relationships between experimental conditions; the inherent variability of the measures, e.g., differences among study participants; and variation in the structure of the experiment conditions, e.g., blocking and order effects. This complexity acts as a major barrier to performing power analysis [Lipsey, 2009, Murphy et al., 2014].

Studies in the natural sciences can rely on meta-analyses of multiple replication studies to suggest effect and sample sizes. However, in VIS and HCI, such replications are rare [Hornbæk and Law, 2007, Kosara and Haroz, 2018] and not highly valued [Greenberg and Bux-

ton, 2008]. Sample sizes (N) are often chosen based on rules of thumb e.g., $N \geq 12$ [Eiselmayer et al., 2019], or drawn from small numbers of studies [Caine, 2016, Hwang and Salvendy, 2010, Hornbæk and Law, 2007]. Studies with human participants also risk *confounding effects* such as fatigue, carry-over, and learning effects. Analytical methods implemented with power analysis tools such as `pwr` [Champely et al., 2018] or G*Power [Faul et al., 2007], are not usually sophisticated enough to account for these effects. Furthermore, researchers must often weigh the benefit of statistical power against high recruitment costs, overly long experiment duration, and the inconvenience of switching between experiment conditions [Mackay et al., 2007]. Although several interactive tools help explore trade-offs among plausible experiment design configurations [Eiselmayer et al., 2019, Mackay et al., 2007, Meng et al., 2017], few address the complex relationship between statistical power and relevant experiment parameters.

Existing power analysis tools are designed as calculators: The user specifies acceptable Type I and Type II error rates, test statistics, experimental design, and an approximate size of the effect. The tool then produces either a single sample size or a chart showing how statistical power increases in conjunction with the sample size, at several effect sizes. We argue that researchers need tools for exploring possible trade-offs between statistical power and the costs of other experimental parameters, especially when the effect size is uncertain.

We propose *Argus*, an interactive tool for exploring the relationship between sample size and statistical power, given particular configurations of the experimental design. Users can estimate parameters – effect sizes, confounding effects, the number of replications, and the number of participants – and see how they influence statistical power and the likely results in an interactive data simulation.

Contributions: We identify challenges and analyze the tasks involved in *a priori* power analysis. We propose *Argus*—which combines interactive visualization and simulation to aid exploration and decision-making in experiment design and power analysis. To demonstrate its efficacy, we describe a use case and a think-aloud study.

3.2 Background and Task Analysis

When planning an experiment, researchers use a strategy called ***a priori* power analysis**[1] to choose which sample size will allow the experiment to detect an expected effect. Power analysis uses the relationship between the **sample size** and the following parameters:

> $\boldsymbol{\alpha}$ is the probability of detecting an effect from an experiment when it is actually absent in the population (Type I error: false alarm). Researchers usually set α based on the convention of each academic field, typically .05 for VIS, HCI, psychology, and the social sciences.
>
> $\mathbf{1-\beta}$, or statistical power, is the probability that a long run of experiments will successfully detect an effect that is true in the population. (β is the probability of a Type II Error: missing the true effect.) If no existing basis exists, Cohen proposed a convention of 0.8 [Cohen, 1988, p.56].
>
> **Effect size** is the difference across means calculated from data under each condition. Researchers make an educated guess of the effect size based on previous research or their experience. Effect sizes are standardized for the calculation, as described in C3 below.

The sample size can be calculated with these parameters, either with software or from a statistics textbook, e.g., [Cohen, 1988]. When facing resource constraints, such as personpower, time or budget, researchers sometimes sacrifice statistical power in exchange for a more attainable sample size. In cases where access to participants is limited e.g., patients, children or other special populations, power analysis may be skipped altogether. Even if the power analysis suggests an unrealistic sample size, it might still offer a useful cost-benefit assessment. In any case, researchers who choose to conduct a power analysis still face the following challenges:

C1: Estimating a reasonable effect size is difficult. Researchers who wish to estimate the effect size face a paradox: The goal of conduct-

[1] Although one can calculate achieved power from data collected during an experiment, such post-hoc analysis is impractical for planning experiments or interpreting the results [Cairns, 2019, p. 110] and [Yatani, 2016, section 5.9.4]. This paper thus uses the term 'power analysis' to refer to *a priori* power analysis.

ing the experiment is to discover the true effect size in the population, but selecting the correct sample size for revealing that effect requires an initial estimate of the effect size. Overestimating the effect size often leads to a sample size that exceeds available resources. Even for studies that can easily scale up the sample size, using an overly large sample size is "wasteful" and an "unethical" use of study participants' time [Button et al., 2013]. Although researchers can conduct pilot studies, finding a large effect size in a pilot with few participants may be misleading and result in an underpowered final experiment [Lakens and Evers, 2014, p. 280]. Cohen proposed a guideline for standardized effect sizes derived from data on human heights and intelligence quotients [Cohen, 1977]. However, reviews in domains such as software engineering [Kampenes et al., 2007] found that the distribution of effect sizes from experiments differ from Cohen's guideline. Therefore, many researchers recommend against using guidelines that are not specific to the domain of study [Lenth, 2001, Cummings, 2011, Baguley, 2004]. In fields where replication studies are scarce, e.g., VIS and HCI [Kosara and Haroz, 2018, Hornbæk et al., 2014], researchers must generate possible effect-size scenarios.

C2: Comparing power at multiple effect size scenarios is necessary. Instead of estimating a single value for the effect size, some researchers estimate the upper-bound—to represent the best case—and the lower-bound—below which the effect is too small to be practically meaningful [Lenth, 2001, Lipsey, 2009, p. 57]—which results in a range of sample sizes to consider (Figure 3.2, A–D). However, in many experiments, the largest attainable sample size may be lower than the one required by the lower-bound effect size (Figure 3.2, C). Researchers must then weigh the benefit of further mitigating risk by increasing the power and the cost of a larger sample size. Because the function between the power and sample size is concave, improving power is increasingly costly [Lakens, 2014b, p. 702] (Figure 3.2, A–B vs. B–C). Among existing software for calculating statistical power, only a few plot the statistical power and the sample size at different effect sizes (see Section 3.3).

C3: Standardized effect sizes are not intuitive. The difference between means is an example of a *simple effect size*, which is based on the original unit of the dependent variable and thus has intuitive meaning for researchers. However, power calculation requires a *standardized effect size*, which is calculated by dividing the simple effect size with a standardizer. The formula for the standardized effect size depends on how the sources of the variances are structured, which in turn de-

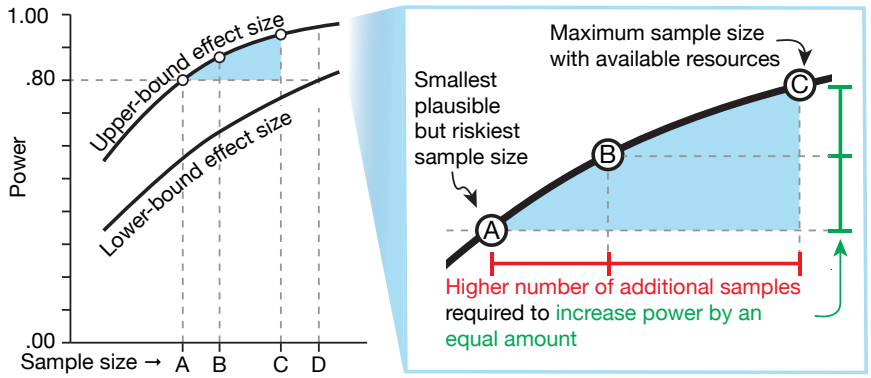


Figure 3.2: Determining power and sample size with effect-size uncertainty and resource constraints.

pends on the experiment design. (See Appendix A for an example on how blocking influences calculation of effect size.) Note how an estimate in the form of a simple effect size may yield different standardized effect sizes. Researchers often have difficulty using standardized effect sizes when choosing their sample size, since these are "not meaningful to non-statisticians" [Baguley, 2004].

C4: Power analysis excludes the temporal aspect of experiment design. Power analysis simplifies sources of variations into a few standard deviations within effect size formulæ. (See Appendix A for an example.) Potential confounds—e.g., the fatigue effect or the practice effect—lose their temporality once encoded into standard deviations. This loss could be a reason that separates power analysis from the rest of the experiment design process [Eiselmayer et al., 2019]. Better integration of temporal effects and design parameters—e.g., number of replications and how conditions are presented to study participants—could allow better exploration of trade-offs.

3.2.1 Task Analysis

Under the What-Why-How framework [Brehmer and Munzner, 2013, Munzner, 2015], the task abstraction could be described as follows. All of the attributes below are quantitative unless stated otherwise.

T1: Come up with an effect size estimate. Simple effect sizes—the difference in the responses between conditions—could have been esti-

mated directly. Alternatively, the estimation can be simplified by first estimating the mean in a baseline experimental condition, and then `deriving` the value of other conditions by `comparing` each with the baseline. The conversion from the simple effect size to the standardized effect size (C3) could be automated when the information about experiment design is available in a computable form.

T2: Check the potential outcome effect size. For experiments with two independent variables or more, the possibilities of the interaction effects could obfuscate how the *a priori* effect sizes influence the final results. (More details in Section 3.4.2) A data simulation could allow the users to `compare` the simulated effect sizes among themselves or to compare them with the specified input—especially in the presence of interaction effects.

T3: Determine candidate sample sizes. Researchers `browse` for the sample size with a reasonable trade-off within a set of constraints (e.g., resources for participant recruitment). To facilitate efficient browsing, they `identify` features of the relationship between power and sample sizes, e.g., where the power-gain is steep or where it plateaus. Multiple scenarios (C2) of effect sizes could also generate different relationships, leading to the need to `compare` their trends.

T4: Try out potential scenarios. Due to uncertainties in effect size estimation (C1), researchers need to be able to `explore` the dependency between their effect size estimates and other parameters—e.g., the fatigue effect (C4)—to the power-sample size relationship. Thus, they need to be able to `record` and review the scenarios. Some changes to the scenarios are categorical—e.g., different choices of counterbalancing strategies. Others are quantitative—e.g., different amounts of the fatigue effect. The abstract data type of the scenarios could be a *multi-dimensional table* with each input parameter as a key and the resulting power as an attribute. However, this abstraction does not capture researchers' exploration traces. Such traces could be abstracted as a *tree* in which each child node is a scenario that is derived based on its parent node.

3.3 Related Work

Before the prevalence of personal computers, researchers used look-up tables [Cohen, 1988, pp. 28–39]) and charts [Scheffe, 1959]) in text-

books to determine the relationship between sample size, effect size, statistical power, and Type I error rate, usually fixed at .05. Early software packages simplified the process by providing command-line or menu interfaces to specify parameters, and displayed a single value for statistical power. Goldstein [1989] surveyed 13 power analysis software packages and highlighted the lack of two key functions: plotting a chart of the trade-offs between parameters, and capturing intermediate results for comparison. Borenstein et al. [1992] pioneered the use of visualization to specify input parameters and inspect relationships among parameters. For input, the tool shows a box plot of the dependent variable by condition on the screen. The simple effect size can be specified by moving the mean and standard deviation of each group with arrow keys or function keys. The software then outputs the effect size and power in real-time. It also produces a chart showing the relationship between power and sample size under multiple effect-size scenarios (see Figure 3.2, left). Nevertheless, due to the low screen resolution, the relationship chart is presented on a separate screen from the input specification, hindering interactive exploration. This tool also restricts analysis to between-subjects designs with two conditions and does not support exploration of the impact of choices in experimental design.

G*Power [Faul and Erdfelder, 2004, Erdfelder et al., 1996, Faul et al., 2007] is one of the most widely used power analysis software tools today. G*Power developers prioritize covering multiple types of statistical tests and high-precision calculation rather then facilitating exploration [Erdfelder et al., 1996]. G*Power calculates power from one set of input parameters at a time. This forces them to record parameters and output at each step of the exploration process. G*Power generates a static chart from a given range of standardized effect sizes.

Some software packages integrate power analysis with experiment design. JMP's design of experiment (DOE) function [SAS Institute Inc., 2016] provides a menu interface for power calculation and generates static charts similar to those of G*Power. The R package `skpr` [Morgan-Wall and Khoury, 2018] provides a menu-based interface for generating experiment designs. However, it only calculates and shows a single power estimate at a time. To explore different effect size scenarios, users must manually save and restore states via their web browser's bookmark function. `skpr` provides a menu interface for generating experiment trial tables and calculating power. However, it provides only the power of the entire experiment design: all variables that take part in the counterbalancing contributes to the power

analysis. Touchstone2 [Eiselmayer et al., 2019] provides a direct manipulation interface for specifying experiment design and displays an interactive chart that visualizes the relationship between the number of participants and power. Unlike `skpr`, users can select a subset of independent variables to include in the power calculation. This lets researchers include nuisance variables in the counterbalancing design, without affecting power calculation. Even so, Touchstone2 does not include confounding effects and relies on menus to specify effect size.

Several researchers have shown that graphical user interfaces (GUI) are better than menus for specifying estimations. Goldstein and Rothschild [2014] compared numerical and graphical interfaces to elicit laypeople's intuitions about the probability distributions of events. They show that users achieve greater accuracy when they can specify distributions graphically. Hullman et al. [2018] support these results in the context of estimating effect sizes for experiments. We argue that power analysis software would benefit from such graphical representations of relationships among parameters, with a GUI to manipulate them.

3.4 Argus User Interface Design

The *Argus* interface is organized into: parameter specification (A–E), simulation output (F–G), and the history view (H) (Figure 3.3). Users begin by specifying metadata about the independent variables in a pop-up window (Section 3.4.1). They can then explore various effect-size scenarios by manipulating the means of the dependent variables for each condition (A). They can also estimate potential confounds (B); and explore how different experiment designs (C–E) influence the outcome (F–G). The history view (H) automatically saves the exploration process and lets users re-load previous scenarios. The rest of this section describes the interface using the example of a 2×2 experiment on how MEDIUM (PAPER v.s. SCREEN) and LAYOUT (ONE_COLUMN v.s. TWO_COLUMN) influences READINGTIME.

3.4.1 Metadata

To facilitate interpretation of simple effect sizes (C3), *Argus* needs the semantics of the dependent variables. Researchers supply this infor-

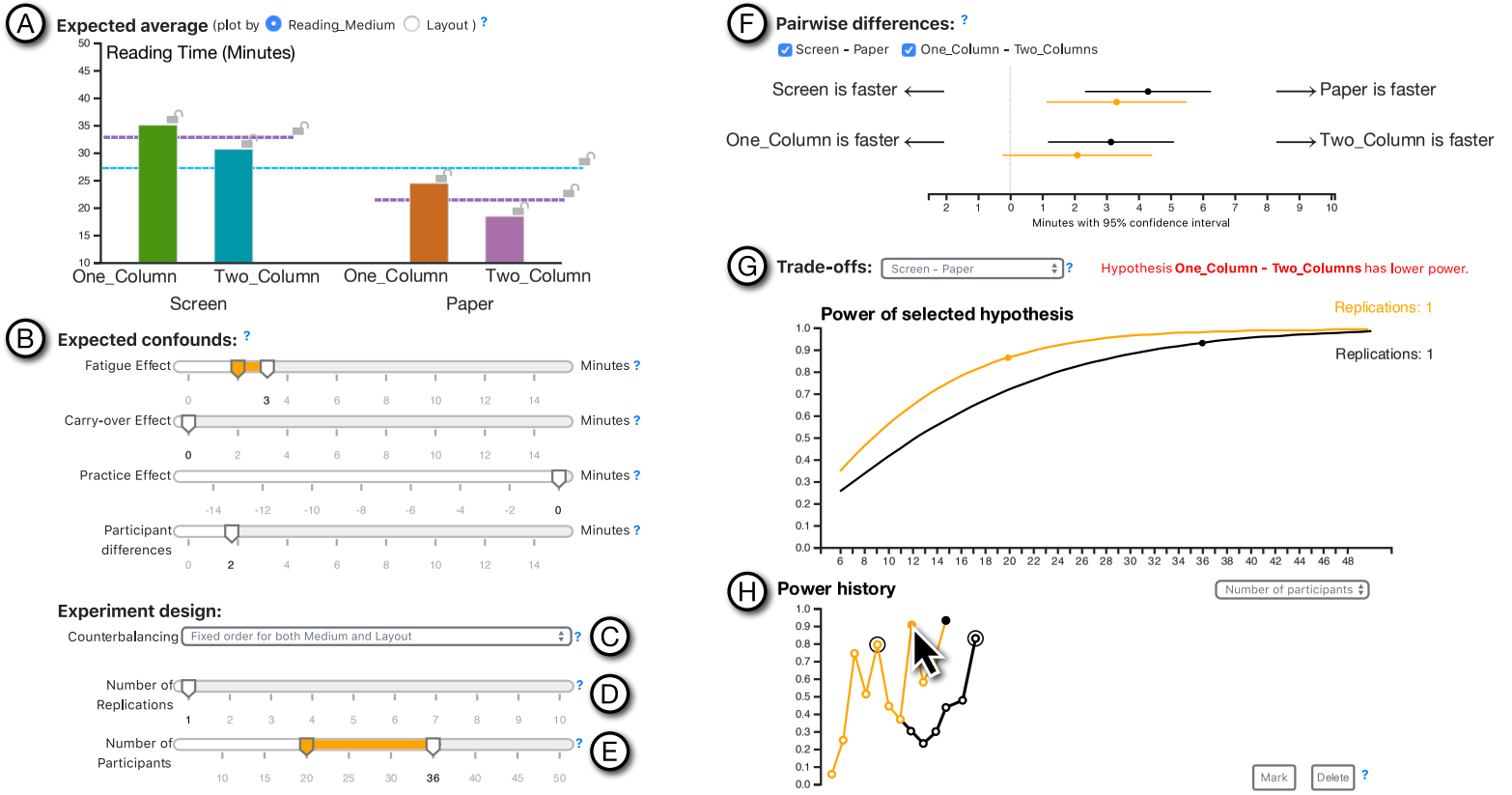


Figure 3.3: *Argus* interface: (Left:) Users estimate effect size by specifying: (A) the expected average for each condition; (B) the relevant confounding effects, and (C–E) the experimental design elements. (Right:) The simulation output includes: (F) pairwise differences, with expected results shown as differences between means; (G) the relationship between power and sample size for making trade-off decisions; and (H) the history view with automatically saved parameter changes. Hovering the mouse over a historical point reveals its settings and results (in orange).

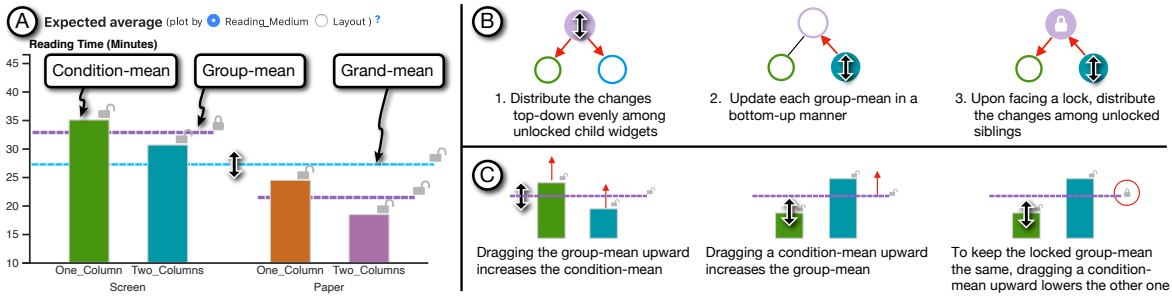


Figure 3.4: (A) Expected average view: users estimate the mean for each experiment condition; (B) Users can lock some means and move others, propagating changes to children, updating group means, or distributing changes to unlocked siblings (no propagation of changes when both the parent and the sibling are locked); (C) Scenarios show: increasing the condition-mean, increasing the group mean, and locking the group-mean.

mation once, at the start of the session. Note that, since many domains use a common set of dependent variables, such as time and error for VIS and HCI, in future, we expect researchers to select relevant dependent variables retrieved automatically from a public domain ontology. Similar ontologies already exist in bioinformatics [Soldatova and King, 2006], and Papadopoulos et al. [2016] have proposed an ontology that specifies dependent variables for VIS and HCI. The current metadata interface is thus a makeshift.

Argus requests the name, unit, expected range, interpretation, and the variability of each dependent variable (DV). *Argus* computes initial ranges for both axes of the interactive charts (Section 3.4.2), and the sliders that adjust various confounds (Section 3.4.4). *Argus* uses the natural-language interpretation, e.g., "30 minutes is *faster* than 50 minutes", to make it easier to read the pairwise plot (Section 3.4.3).

3.4.2 *Expected-averages* View

Argus uses a direct manipulation interface to determine effect sizes, which lets users work with simple effect sizes (T1) and explore multiple effect-size scenarios. Instead of specifying mean differences, *Argus* lets users specify the expected mean of each experimental condition. This condition-mean specification lowers user's cognitive load because they can flexibly estimate each condition individually.

Argus presents the condition-mean relationship as a bar chart (Figure 3.3.A), and the bar colors are drawn from the 2D colormap of Bremm et al. [2011] by assigning one dimension per variable[2]. Users can estimate each condition-mean by dragging the bar vertically. Horizontal lines encode the *group-mean*—calculated from all conditions of an independent variable—and the *grand-mean*—calculated from all independent variables (Figure 3.4.left). Despite the potential for within-the-bar bias [Correll and Gleicher, 2014], encoding the bars keeps condition-mean visually distinct from the group-means and the grand-mean. Users can switch the hierarchy level of the condition axis in the bar chart via radio buttons. We describe two common use cases for expressing effect size:

Main effects occur when a particular level of an independent variable causes the same change in the dependent variable, regardless of the level of other independent variables. For example, a main effect of MEDIUM on READINGTIME could be that reading on a SCREEN is generally slower than reading on PAPER. To specify this as a main effect, the user would have to drag two bars (ONE_COLUMN and TWO_COLUMN of the SCREEN condition) upward by equivalent amounts. This becomes tedious when the independent variable has many levels.

Interaction effects occur when the mean within each group differs according to the level of another independent variable. Suppose we want to express how the LAYOUT affects READINGTIME. As above, we register MEDIUM as a main effect, but ensure that the group means for SCREEN and PAPER remain the same.

If the user changes the ⟨ONE_COLUMN, SCREEN⟩ bar, the group-mean of the SCREEN condition will also change. To keep the same group mean, the user must first remember the group-mean prior, and then adjust the other bars to compensate.

Both scenarios involve manipulating multiple conditions simultaneously by dragging group-means and the grand-mean. Users can also lock some means while changing the rest, and the system automatically propagates the changes. However, enabling this interaction technique is tricky because of the hierarchical dependency among these values.

[2]We use the Color2D library: dominikjaeckle.com/projects/color2d/

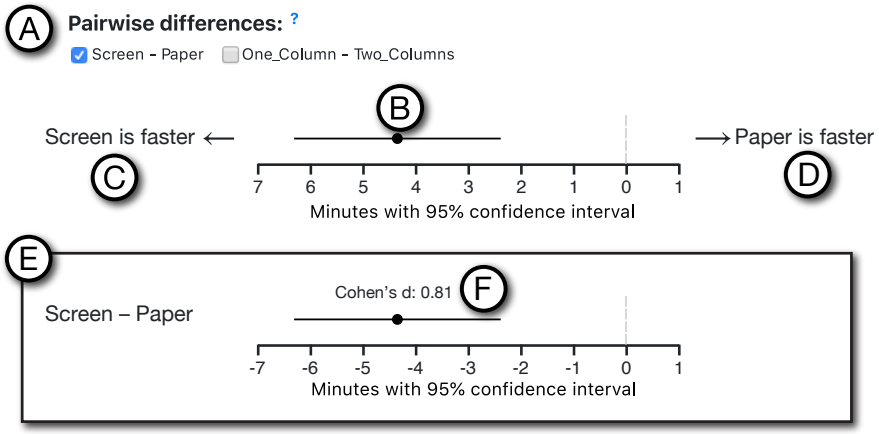


Figure 3.5: (A) Pairwise-difference view for selecting which effects to include. (B) Dancing confidence interval shows the mean differences, with (C–D) natural language labels on either side. (E) Holding a Shift key displays labels for mean difference and Cohen's d (F).

Argus implements a propagation algorithm (Appendix B and Figure 3.4, right). The relationship between the hierarchy of means is represented as a tree rooted at the grand-mean. A change to a parent node—the grand-mean—is first recursively propagated to the children, e.g., group-means and then the condition-mean. The amount of change is distributed evenly to all unlocked children. After finishing the change propagation, the update moves upward. If the update reaches a locked parent, the change is distributed to any unlocked siblings. The propagation algorithm offers users flexibility, letting them switch seamlessly through different representations at different levels, not only individual conditions, but also main and interaction effects.

3.4.3 *Pairwise-difference* View

To help users evaluate the consequences of their effect size estimates (T2), we simulate the data and show the difference between means and their confidence intervals in the *Pairwise-difference* view (Figure 3.5). The horizontal axis shows the difference in the original unit of the dependent variable—a simple effect size (C3). The horizontal axis lists all possible comparison pairs. An independent variable with m levels can accommodate $\binom{m}{2}$ pairwise comparisons. For each pair, we show the mean difference, displayed as a black dot, together with

its 95% confidence interval, displayed as a black line. Unlike the bar charts used for input (Section 3.4.2) this reduces bias [Correll and Gleicher, 2014]. Although violin plots reduce bias somewhat, we chose the dot-and-line display because they can fit more lines into a limited space. This is crucial when comparing two sets of parameters side-by-side with the history function (Section 3.4.4).

In Figure 3.5.B, the difference appears to the left of the zero indicator. Had we presented the result on a normal number line, it would have appeared on the negative side, and the chart could have been interpreted as: "the difference is around minus 4 minutes". Since reading double negatives is cognitively demanding, we present absolute values on both sides of zero on the horizontal axis, and add annotations on the left and the right margin (C and D). This makes it easier for users to interpret, e.g., "SCREEN is faster for around 4 minutes". Users can press-and-hold the shift key to show the normal number line with negative values on the left of the zero, in Figure 3.3.E. This mode lets users change the label on the left margin to present a mathematical difference ("SCREEN- PAPER"). For advanced users, *Argus* also annotates Cohen's d standardized effect size above each confidence interval.

In Figure 3.3.F, both SCREEN-PAPER and ONE_COLUMN-TWO_COLUMN are selected. Suppose we are only interested in comparing reading media because the layouts were included as a nuisance variable. Deselecting the "ONE_COLUMN- TWO_COLUMN" checkbox might yield a slightly narrower confidence interval for the "SCREEN- PAPER" difference. The reason for this improvement is that the difference between the two layouts is slightly smaller in the PAPER condition (Figure 3.3.A), i.e. there is an interaction effect.

Since *Argus* shows simulated data instead of real data collected from an experiment, we need to ensure that users are aware of the uncertainty generated by the simulation. We thus use the *dance of the CIs*, a time-multiplexing approach that shows the results of multiple simulations in the same figure [Cumming, 2012, Dragicevic et al., 2019]. The animation runs in 2 fps, to allow the user to notice changes between frames [Trick and Pylyshyn, 1994]. An alternative to the dance animation is a forest plot that displays all confidence intervals from the simulation next to each other, with a diamond shape to summarize them [Cumming and Calin-Jageman, 2017, Chapter 9].

We chose the dance because it uses less screen space, and motion is a strong visual cue. Even when the user focuses somewhere else on the

screen, the animation is registered in their peripheral vision. In addition, users can pause the animation and navigate individual frames by the left and right arrow keys on the keyboard.

3.4.4 Exploring Trade-offs

At each effect-size scenario, users can increase power by adding more participants, increase the number of trial replications in the counterbalancing design, or both. Some experiments may be constrained by participant fatigue and need to limit the duration, whereas for other experiments, the cost of recruiting additional participants may outweigh the drawbacks from the fatigue effect. Argus lets users explore how different experiment design scenarios and confounds can influence power (T4), as shown in Figure 3.3. Users estimate levels for each potential confounding effect (B) and select an experiment design parameter accordingly (C–E). They explore how the trade offs change based on sample size and power (G), and can revisit and compare earlier explorations with the *History* view (H).

Confound Sliders

Confounding effects can be specified by sliders (Figure 3.3.B). When users drag a confound slider, *Argus* shows a pop-up overlay to preview its effect (Figure 3.6). The pop-up is a bar chart showing how the measurement of the dependent variable (vertical axis) could change along with the experiment trials (horizontal axis). The order of trials and the effects are calculated based on the choices in the *Experiment-design* view (Section 3.4.4).

Four types of confounds are of interest in power analysis [Lazar et al., 2017b]. For readability, we will explain each of them in terms of reading time. Increasing the *fatigue effect* (Figure 3.6.A) would cumulatively increase the reading time for each subsequent trial (Figure 3.6.B). The *carry-over effect* (Figure 3.6.C) occurs when the user is unfamiliar with the task itself: Their performance is worst in the first trial, but gradually improves over subsequent trials, regardless of the experimental condition. The practice effect has two variations: The *within-condition practice effect* (Figure 3.6.D) represents improvements resulting from the participants' familiarity with each experimental condition. Thus, improvement in one condition does not influence

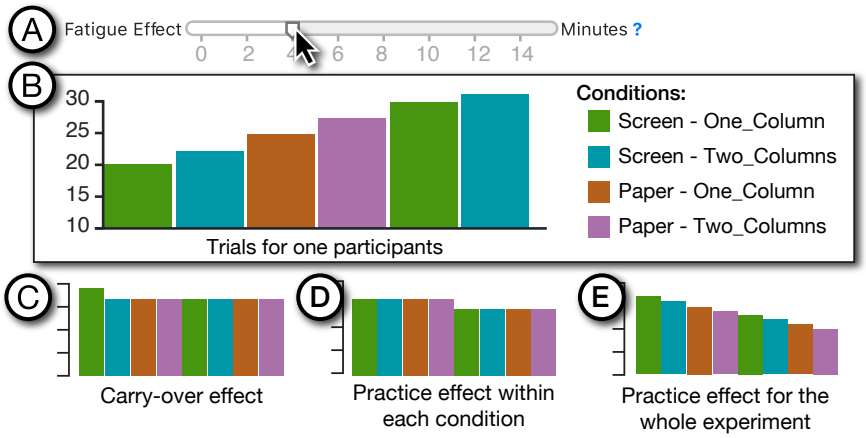


Figure 3.6: (A) Adjusting the 'fatigue' confound effect level (B) displays its corresponding influence on the data, as well as (C) carry-over effects, (D) practice effects per condition and (E) for the whole experiment.

subsequent trials in other conditions. The *whole-experiment practice effect* (Figure 3.6.E) results from users' familiarity with the task, regardless of experimental condition. This is the opposite of the fatigue effect. A participant in our think-aloud study (Appendix D) pointed out the difference between these two practice effects, and we plan to incorporate the whole-experiment practice effect in the next version of *Argus*.

The confound pop-ups use a bar chart to encode the level of the dependent variable. We take advantage of the Gestalt law of similarity to let the user associate the color-coding of conditions to those in the *Expected-averages* view. Future versions of *Argus* could include a more advanced interaction technique that lets users specify a range or a probability distribution for each confounding variable.

Argus uses the dependent variable metadata (Section 3.4.1) to determine the range for each slider. The direction of the available values depends upon which direction users specify as the "better" direction. For example, in Figure 3.3.B, the variability is set to ± 5 minutes, and the interpretation is specified as "slower is better". These settings create a fatigue-effect slider ranging from 0–15, and a practice-effect slider ranging from -15–0. All sliders are initially set to zero to represent no confounding effects. *Argus* also provides an additional slider for specifying variations across participants.

Experiment-design View

The effect of confounds such as the fatigue effect could even out across participants if the experiment is properly counterbalanced. In the running example, the experiment has four conditions. A complete counterbalancing would require covering the $4! = 24$ possible orderings of the conditions, which would in turn require recruiting a *multiple* of 24 participants. Alternatively, users might consider using a standard Latin Square design, which addresses the order effect between adjacent trials. This Latin Square design requires only multiples of four participants, allowing for greater flexibility in the sample size.

Recruiting fewer participants than required multiple may lead to an imbalanced experiment, and affect both the observed effect and power. Finally, users could collect several replications of data from each participant. This number of replications influences the trial table, and thus influences how the confounding effects contribute to the data.

In the field of HCI, several tools exist for counterbalancing design [Eiselmayer et al., 2019, Mackay et al., 2007, Meng et al., 2017]. Eiselmayer et al. [2019]'s interview study suggests that counterbalancing design and power analysis are performed in two separate loops. We envision that users should use one of these tools to come up with experiment design candidates. Then, these candidates can be imported to *Argus*. For these reasons, we present a minimal user interface for counterbalancing design: a drop down list for selecting the counterbalancing strategy (Figure 3.3.C) and two sliders for the number of replications (D) and the number of participants (E). These controls work together with the *Power Trade-off* view and *History* view.

Power Trade-off View

The *Power Trade-off* view (Figure 3.3.G) is the heart of power exploration (T3). It visualizes the outcome of the adjustments in *Expected-averages* view, *Confound* sliders, and *Experiment-design* view. The visual encoding is based on the chart relating power vs. sample size, commonly used in statistics textbooks, e.g., [Scheffe, 1959]. The sample size appears on the horizontal axis and the power on the vertical axis. The current selection of the sample size is represented as a dot, and the relationship between these two parameters are displayed as a

black curve. We used this encoding despite the fact that the underlying data is discrete—the sample sizes are integer—because curves facilitate interpretation of the local rate of change Cleveland and McGill [1986], which is usually the case when researchers assess power trade-offs.

Touchstone2 [Eiselmayer et al., 2019] enhanced this textbook chart by automatically showing the confidence band around the current parameter set, which was calculated from a single "margin" parameter. In *Argus*, variations in power can originate from any of a combination of multiple sources, e.g., effect size or confounds, making it difficult to determine which are associated with the confidence band.

Argus enhances this chart in two ways: First, Users can switch the horizontal axis between the sample size and the number of replications. Setting the axis to the sample size shows the number of replications annotated on the right end of the power curve. This switch could be used when the sample size faces a stricter constraint than the number of replications, or vice versa. In Figure 3.3(G), suppose the resource constraint allows the recruitment of a maximum of 24 participants, which results in the power of 0.7. Users can now consider the trade-off between the number of replications and power.

Second, *Argus* shows the chart individually for each of the pairs of independent variable levels, e.g., Figure 3.3.G, shows "SCREEN- PAPER"). Users can change the pair with a drop-down menu. *Argus* shows a warning if any pairs produce lower power than the current pair. The user can also select the "Minimum power" option to always display the pair with the lowest power. Although this pair-selection is also present in the *Pairwise-difference* view, the selection in *Power Trade-off* view is independent: Switching it does not trigger a simulation. This independence allows the user to explore nuisance factors without changing how the confidence interval of differences is calculated.

***History* View**

The *History* view (Figure 3.3.H) ties together all above-mentioned views to enable exploration of scenarios in light of uncertainty from effect size estimation and confounds (T4). *Argus* thus improves on

other power analysis systems that force users to record each scenario's output before manually comparing them. (Section 3.3).

Each step of parameter adjustment is recorded automatically in an abstract tree. The root of the tree is the initial setting of zero effect size with no confounding variables. The tree is visualized on a two-dimensional cartesian coordinate with the vertical axis showing the power. The horizontal axis shows the depth of the node from the root. Each node is encoded as a white circle with black outline, and it is connected to its parent node with a line. The current node is encoded in a black circle to associate it to the the dot in the *Power Trade-off* view with the Gestalt principle of similarity. Adjusting a widgets in the views mentioned above creates a child node. Clicking on a past node restores its parameters all other views. The restoration excludes the selections in the *Power Trade-off* view to enable users to retain their current focus, as described in Section 3.4.4. During exploration, it is likely that only a few nodes will be of interest. Users can mark/unmark a node by clicking a button. An additional concentric outline circle is added to each of the marked nodes.

In addition to restoring the parameters, users may hover their mouse cursor over a node to preview its parameters and output. The preview values are shown in orange, simultaneously with the values of the current node in black (Figure 3.3). We use juxtaposition and superposition faceting techniques. These two techniques were analyzed in Javed and Elmqvist [2012]'s survey of composite visualization. Their analysis found that for tasks that focus on direct comparison in the same visual space, superposition is more effective than juxtaposition. For the *Power Trade-off* view, since decisions about sample size usually take place around the few crucial values (see C2 and Figure 3.2), we superpose the curves. For the *Confound* sliders and *Experiment-design* view, the sliders and the drop-down list, preview values are also superposed. For the *Expected-averages* view, however, both superposition and juxtaposition would be appropriate. Here, superposition allows the bars representing the current state to provide a stable visual anchor.

For the *Pairwise-difference* view, the uncertainty communicated by the animation would be muddled when two superposed confidence intervals overlap. Therefore, we juxtapose the preview error bars side-by-side (Figure 3.3.F). For the *History* view itself, we highlight nodes and edges in the current branch during preview.

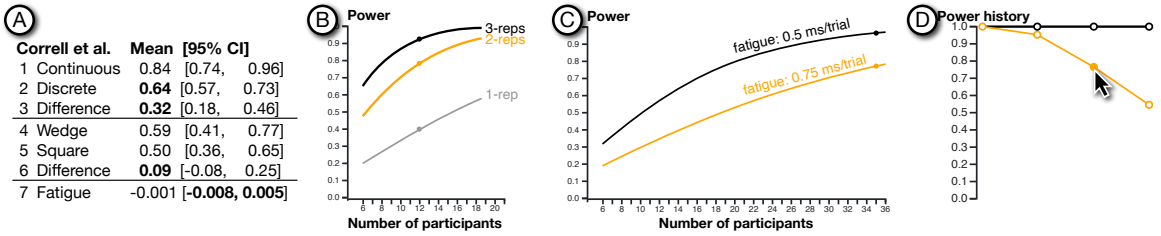


Figure 3.7: (A) Relevant error estimates based on Correll et al.'s data; (B) The power is plotted against the number of participants 1-, 2-, and 3-replication scenarios. (In Argus UI, only the maximum of two curves are shown at a time during interactive comparison.) (C) Power trade-off curve of three-replication with the fatigue effect of 5 ms (in black) and 7.5 ms (in orange). (D) The History view showing two branches: three-replication (in orange) and two-replication (in black).

We also decided to limit the comparison to two nodes—the current node and the preview node—to reduce visual complexity. A pairwise comparison of historical nodes together with the marking functions allows users to gradually narrow down the parameter choices.

Scaling the Design for More Complex Experiments

Our prototype supports within-participants designs with two independent variables. More complex experiment designs may have more than two independent variables, and each independent variable could have more levels. Only two views will be affected: The *Expected-averages* view could present more levels by incorporating the fish-eye technique [Rao and Card, 1994]. To address more independent variables, the system should allow the users to reorder the hierarchy in the horizontal axis—e.g., by drag-and-drop. Users should also be able to exclude some of the independent variables from the axis, which will summarize several bars of the same level into one, which further reduces the visual complexity. As for the *Pairwise-difference* view, scrolling and panning could be necessary to handle the increased number of pairs. When their effect sizes are very different in the magnitude or sign, the comparison could be broken down into subsets, presented in separate windows.

3.5 Implementation Details

Argus was written in HTML and JavaScript. We used D3.js[3] for interactive visualizations. Experiment designs are implemented in the TSL language and trial tables are generated on the client-side with the TSL compiler [Eiselmayer et al., 2019]. Statistical calculations are implemented in R[4], and Shiny[5]. We used a MacBook Pro (2.5GHz, 16GB memory, MacOS 10.14) for all benchmark response times.

To enable interactive exploration in *Argus*, we make the following three implementation details that differs from standard statistical procedure for *a priori* power analysis and post-study statistical analysis.

3.5.1 Monte Carlo Data Simulation

Power can be calculated from an α probability value, a standardized effect size, and a sample size. However, incorporating confounds, e.g., a fatigue effect, is analytically complex (C4). Instead, we use a Monte Carlo simulation, based on algorithm 1 of [Zhang, 2014]: First, a population model is created programmatically, based on an estimate of the mean and the standard deviation (SD) of each condition. From this population, we sample data sets and use them to calculate statistics. The Monte Carlo paradigm has been shown to be robust for tricky cases such as data that are not normally distributed, missing data, or mixed distributions [Muthén and Muthén, 2002, Schoemann et al., 2014, Zhang, 2014].

We extend the algorithm to incorporate confounding variables: First, we obtain a trial table for the specified experiment design from the TSL compiler. Based on the trial table's structure, we generate each confounding effect specified by the user in the interface (Section 3.4.4). For example, a two-second fatigue effect for movement time cumulatively lengthens each subsequent trial by two seconds. All confounding effects are added to each simulated data set before calculating statistics. Data simulation and confounding calculations are vectorized. On average, we can generate a data set with 50 participants and

[3] d3js.org
[4] r-project.org
[5] shiny.rstudio.com

10 replications with all confounding effects in place, in less than 30 ms on our benchmark machine.

3.5.2 Making Power Calculation Responsive

Calculating statistical power is computationally expensive because it requires a numerical integration between two overlapping probability distributions (see Fig. 11 of [Eiselmayer et al., 2019]). Furthermore, post-hoc power calculation uses an *observed effects size* from the data, which may differ from the input effect size due to confounding effects. To calculate observed effect sizes, we must fit a general linear model for each data set. In normal statistical analysis, such model-fitting is done only once, so results appear almost instantaneously. However, plotting the chart of sample size and power (Figure 3.2) requires one calculation per simulated data set. By default, *Argus* generates 1000 data sets for each sample size. Here, we show the sample size from 6 to 50. On our benchmark machine, the entire calculation takes around two–three minutes.

To ensure the responsiveness of the user interface, we first approximate the observed effect size with a pairwise Cohen's d calculated with the `pwr.t.test` function from the `pwr` package [Champely et al., 2018]. The average turn-around time is 200 ms. Model-fitting results are sent progressively to the user interface, which updates accordingly. We further ensure responsiveness, we also make further tweaks in the communication between R, Shiny, and Javascript as detailed in Appendix C.

3.5.3 Statistical Model and Pairwise Difference Calculation

After modeling participants as a random intercept, we derive the observed effect size and the pairwise difference in terms of means and confidence intervals from mixed-effect models. (See Fry et al. [2016]'s HCI statistics textbook for more details on the model choice.) *Argus* automatically formulates a mixed-effect model and a contrast matrix for generalized linear hypothesis testing, based on the user's choice of the condition pairs of interest (Section 3.4.3), We use the `lme4` package [Bates et al., 2015] for model fitting and the `multcomp` R package [Hothorn et al., 2008] for the test. Confidence intervals are calculated

with a single-step adjustment with the family-wise error rate set at $\alpha = .05$.

3.6 Use Case

To demonstrate how to use Argus, we draw an example from a study on color ramps from Smart et al. [2020]—of which the study plan could have been informed by a similar study by Correll et al. [2018]. Additionally, both studies made their data publicly available, allowing us to derive additional information for planning and testing. We first describe the background of both studies—which constrains the parameter space to be later explored with Argus. To aid cross-referencing, we highlight relevant values in **bold**. Calculation details are provided with R code in supplementary S2.

3.6.1 Background

Smart et al. [2020] propose to generate color ramps based on a corpus of expert-designed ramps by using Bayesian-curve clustering and k-means clustering. Their experiment compared four types of ramps (BAYESIAN, K-MEANS, DESIGNER, and the baseline LINEAR) in three visualization types (scatterplots, heatmaps, choropleth maps), in a total of 12 conditions. In each experimental trial, study participants are asked to identify a mark on the visualization that matches a given numerical value. They measured errors and aesthetic ratings. Because a comparable aesthetic data were unavailable in prior works, this use case focus only on the errors, which is defined as $|v_{\text{given}} - v_{\text{selected}}|$.

To plan their study, Smart et al.'s study could have leverage information from Correll et al.'s experiment [6]. The latter used the same identification task, albeit only heatmaps are used as the visualization. Their study investigated how color ramps can be used to encode both values and uncertainty. Although their experiments have different conditions compared to Smart et al.'s, two of their results are relevant: (1) the significant difference between continuous vs. discrete color map, and (2) the absence of a statistically significant difference

[6] Although Smart et al. [2020] mentioned that their study was similar to [Gramazio et al., 2017], the latter concerns categorical palettes rather than quantitative color maps.

between wedge-shaped vs. square-shaped color legend. The former can be used as an upper-bound and the latter as a lower-bound for the effect sizes. Since Correll et al.'s accuracy was defined differently from Smart et al.'s error, we use Correll et al. [2018]'s data to calculate the errors—which result in the statistics shown in Figure 3.7.A.

In addition to the effect sizes, we also retrieved the duration information. In each trial of the relevant experimental condition, participants took 8.5 seconds. Since the stimuli of Smart et al.'s study was four times larger, we extrapolate **each trial to take 34 seconds**. In Correll et al. [2018]'s study, the median session duration was 13.5 minutes. We also analyzed the data for the fatigue effect and found it negligible with the estimate in Figure 3.7.A, row 7.

Smart et al. [2020] recruited **35 expert designers** as their study participants; we use this number as a maximum number of participants. On the opposite, we consider **12 as a minimum number of participants** based on a rule of thumb [Eiselmayer et al., 2019]. Since the participants were experts, they might be less willing to participate in a long study. Therefore, we constrained the longest session duration to 30 minutes. Leaving 5 minutes aside for instruction and informed consent, this results in **the maximum of 3 replications** ((25 minutes $\times$ 60 seconds) $\div$ (12 conditions $\times$ 34 seconds) = 3.6, rounding down) We used the randomized counterbalancing according to Correll et al. [2018]'s design. We will aim for power above 0.8—according to Cohen's recommendation [Cohen, 1988, p. 56].

3.6.2 *A priori* Power Analysis

In the following scenario, the goal of the researcher[7] is to determine the sample size (number of replications and number of participants) for his experiment. As mentioned above, these decisions are constrained by the total duration of the session, maximum number of participants, and potential for confounding effects. The exploration starts with the upper-bound and lower-bound scenarios and proceeds to explore a potential fatigue effect.

[7] The researcher will be further referred to as a gender-neutral "he".

Upper-bound Scenario

He started with 12 participants and 1 replication. He moves the grand mean to 0.64 and the group-means of conditions other than the LINEAR to 0.32 **(T1)**. These values are from Correll et al. [2018] discrete conditions (Figure 3.7.A, row 1), and its difference to the continuous conditions (Figure 3.7.A, row 3). On the Power Trade-off view, the researcher sees that the power of the effect between LINEAR– DESIGNER pair almost 1.0, which is very high—indicating that if the effect size is large, only 12 participants would be adequate **(T3)**.

Lower-bound Scenarios

He moved the group-mean of the DESIGNER condition to 0.55 (from Figure 3.7.A, row 6). The power drops to around 0.4. One way to address this is to increase the number of replications to 2 and 3, resulting in the power of 0.7 and 0.9 respectively **(T3)**. He hovers his mouse cursor on the history nodes to superpose the power curves in Power Trade-off trends (Figure 3.7.B). According to the curve, for one- and two-replication designs, adding participants would dramatically increase power. However, for 3-replication setting already have relatively high power **(T3)**.

Naturally, the researcher would hope that the BAYESIAN and K-MEANS will be better than DESIGNER ones. However, he does not know *a priori* which of the two algorithmically-generated ramps will be better. To reflect these beliefs, he moved both BAYESIAN and K-MEANS to 0.46 **(T1)**. These values reflect a small effect when comparing with DESIGNER condition. However, when comparing with LINEAR condition, the difference is sizable. In the Power Trade-off view, he switches to the pair Designer – Bayesian and found the power to be above 0.8 **(T3)**. The pair-wise difference (Figure 3.8) shows the difference between all pairs except BAYESIAN vs. K-MEANS to be larger than zero. Also, the difference between LINEAR and the two algorithmic conditions is larger than between LINEAR and DESIGNER. Results like these matches the researcher's expectation; therefore, he marked this point in the History view as a plausible design **(T2)**.

Fatigue Effect Scenarios

From the scenario above, the total duration of a study session is 20.4 minutes (3 replications $\times$ 12 conditions $\times$ 34 seconds/trial). This duration is longer than Correll et al. [2018]'s median of 13.5 minutes. Therefore, it is possible that the fatigue effect may have influenced the experiment. To explore its impact, he adjusts the fatigue effect to 5, 7.5, and 10 ms per trial—according to Figure 3.7.A, row 7—and found that the power drops very low **(T4)**. Therefore, he changes his exploration strategy to determine how much of the fatigue effect could his study design tolerate at the maximum number of participants of 35.

He set up the 35 participants without any fatigue effect as a starting point and mark it in the History view. Then, he creates two branches of scenarios: two- and three-replications. In each branch, the explore the three levels of fatigue effects mentioned above **(T4)**, resulting in Figure 3.7.D. The two-replication scenarios seem not to change the power much **(T3)**—and hence robust to the fatigue effect. However, collecting two data points per condition could be susceptible to outliers.

On the other hand, in the three-replication branch, the power reduces dramatically as the fatigue effect increases **(T3)**. By selecting one node (fatigue: 5 ms/trial) and hovering on another (fatigue: 7.5 ms/trial), he can compare the two corresponding curves in the Power Trade-off view (Figure 3.7.C). From the orange line in this chart, he can see that if the fatigue effect is higher than 7.5 ms, the experiment will need more than 35 participants to achieve power at least 0.8. He could not effort this scenario **(T3)**.

To decide between the susceptibility to outliers or the fatigue effect, he could run a pilot study to assess the impact of the fatigue effect with the three-replication setting. If the fatigue effect is 0.5 ms/trial or lower, an experiment with only 22 participants would be adequately powerful. We validated this potential choice by a simulation that resamples data from Smart et al. [2020]'s result and found that recruiting only 22 participants are likely to generate similar outcome as those reported in Smart et al.'s paper. The simulation details is provided in supplementary S2.

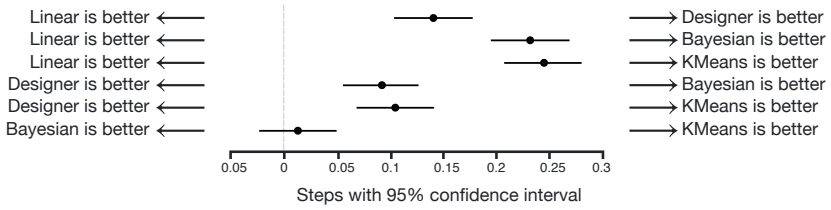


Figure 3.8: The pairwise difference plot from the case study.

3.7 Think-aloud Study

To better understand how *Argus* users could be used in power analysis, we conducted a formative study that aims to answer the following research question: What insights can researchers gain from being able to interactively explore the impact of design choices for their experiments. The study was preregistered (Anonymized URL) and is fully described in Appendix D. This section provides a summary.

3.7.1 Method Summary

Participants

Nine researchers in HCI and/or VIS participated in our study. Five of them were experienced researchers who has conducted three or more experiments. They were either senior scientists (post-doc or higher), and one was a senior-year Ph.D. student. The rest of them were Ph.D. students or post-docs who had learned about experimental method, but had planned less than three experiments. Henceforth, the participants in our study will be referred to as "users" To avoid confusion with the "number of participants" term in Argus.

Task and Procedure

We used a think-aloud protocol where users voice their observations and reasoning [Lewis, 1982]. The users watched a video explaining *Argus* and relevant concepts in experiment design and statistics. Then, they used *Argus* to determine a sample size for a Fitts's law experi-

ment based on a summary of prior findings. At the end of the session, we interviewed and asked them to rate their experience.

Data Analysis

We recorded users' screen and audio think-aloud and interview responses. We performed a qualitative analysis with bottom-up affinity diagramming with the focus on insights [Saraiya et al., 2005].

3.7.2 Selected Results

Overall, the majority of the users reported that they have gained new insights about experiment design:*"the preview is very useful to understand the confound effects."*($P9_N$). $P7_N$, $P8_N$ were not familiar with carryover effect and practice effect but they expressed their understanding of the difference between these effects when they saw the previews. Five users applied their experience in conducting experiment to consider potential confounds. For example, $P8_N$ said *"adding more replications can yield higher power but participants may be tired [so] I need to increase the fatigue."* after increased the number of replications.

The influences of the number of replications and participants to power were explicitly observed: *"The power is very high now. I am going to tweak replications and participants to see how power is going to change [...] reduce the number of participants, power drops down. It makes sense"* (P4). Participants also interpret the characteristics of the curve in *Power Trade-off* view: *"The power get stabled after a certain number of participants. The current number of participant is a bit too much. We can reduce the number"* (P5).

However, three of the expert users were initially puzzled why changing the practice effect slider did not influence the mean-differences nor the power. The study moderator had to point out that the effect was prevented by the Latin-square counterbalancing, or because only one replication was used. This result suggests an opportunity to improve users' awareness when causal links are muted by a moderating parameter. (See the transition matrix in Appendix D for how users inferred the causality between power analysis parameters)

Five users tweaked expected confounds and observe how the power of adjacent nodes in the *History* view gradually changes. Four users repeatedly used the hover function to preview the difference. Two expert users use the branching to explore multiple strands of parameter configurations. These behaviors show that the *History* view successfully facilitates the exploration of statistical power.

3.8 Lessons Learned

We have went through many cycles of design, prototyping, and testing. It was fascinating to see how the context of use (statistics) influence users' expectation and behavior when interacting with *Argus*. We would like to share three lessons:

L1: Enabling visual exploration and close-loop feedback generates curiosity about causal relationships. The *History* view enables users to compare different scenarios. Our task analysis shows that the focus of comparison is the relationship between the statistical power and sample sizes. Therefore, in an early version, hovering the mouse cursor on a historical node showed the differences only in the *Power Trade-off* view and the *Pairwise-difference* view. For other views, the input parameters were temporarily reverted back to the state of the historical node. For example, the knob of confound sliders is positioned at the state of the historical node. However, users who tested this version of *Argus* are curious to see the differences in the input parameters as well. We surmised that the immediate feedback from simulated data and the the affordance for parameter exploration piqued their curiosity of the causal relationship between each of the input parameter to the power. This evolution of users' need is another evidence that visualization design is essentially iterative.

L2: The ease of verbalization could be important for integrating the domain knowledge to interpret visualized data. In *Pairwise-difference* view, we used points and error bars to visualize the results of simulation. An early version of *Argus* shows output in terms of arithmetical difference (Figure 3.5, E). Some users struggled to understand the effect when the difference falls on the left of the zero. To address this problem, we changed the default display mode to show natural language labels (Section 3.4.3). After this addition, we did not observe this difficulty. Automatically-generated verbal description of visualization has been shown to help users understanding statistical test

procedures [Wacharamanotham et al., 2015] and to support understanding of machine-learning models [Hohman et al., 2019]. We conjecture that, for the tasks that requires users to combine visual interpretation with their domain knowledge, verbalization is important for the users to successfully integrate visual processing with their knowledge.

L3: When asking for a ballpark, avoid precise terms. *Argus* needs a rough approximation of the standard deviation (SD) of the population of the dependent variable to initialize the range of the confound sliders. This initial value is important to set an appropriate range and granularity of the sliders. However, it does not need to be precise. After the sliders are initialized, users can come back to change this value any time to expand or contract the range of the slider. In an earlier version, the UI simply asked the user to input a number into a text field with the label *"Approximated SD"*. This question turned out to be difficult for people we pilot-tested the software with. Some of our colleagues even invested time to lookup research papers in order to give an accurate value. In a later version of *Argus*, we reworded it to *"Variability"*, which is a broader term that could be understood as, e.g., SD, variance, or simply a range. This change seems to lower the users' anxiety and proceed to use *Argus* faster. We conjecture that the context might have also putting the users unnecessarily on guard. Pilot testing with users are helpful to identify such unintended barriers, especially for the choke points of the task flow.

3.9 Discussion

Argus is another addition to the ecology of tools developed in the VIS and HCI community aiming to improve practices in experiment design and statistical analysis. Like previous works [Wacharamanotham et al., 2015, Eiselmayer et al., 2019], *Argus* demonstrates the power of direct manipulation interfaces to assist in the tasks previously dominated by menu- or command-based interfaces. These works add interactivity to existing domain objects (statistical charts and trial tables) to allow the users to *specify*, *compare*, and *explore* diverse outcome possibilities. These common interaction capabilities and the mappings between abstract concepts in experiment design and statistics to interactive visualizations seems to suggest **an emerging design pattern for a more usable software tools for research scientists**.

The challenges that these works—including *Argus*—face is the limited user to participate in evaluation studies. In other words, our studies have low power— while we are advocating for the importance of powerful studies. Specifically, we face **a trade-off between the coverage of use cases (e.g., which experiment designs to support) and realism of the studies.** For *Argus*, we set the scope of use cases by pre-determining the scenarios for the study participants. Although this makes the implementation tractable, the participants might be less motivated to explore—compared to when they design their own experiments. However, researchers usually design and conduct only a few experiments per year, which imposes a challenge of collecting meaningful longitudinal data. On the other hand, one could assess learning achievements by novices (e.g., as in [Wacharamanotham et al., 2015]), but it is unclear how much the design implications drawn from such learning studies could apply to experts. In summary, **we need a methodology that allows studying infrequent knowledge works being conducted by experts**.

3.10 Conclusion

Our goal is to help VIS and HCI researchers consider statistical power when planning their experiments with human participants, which requires performing *a priori* power analysis. This paper provides three key contributions. First, we present a detailed **analysis** of the problems faced by experimenters and identified key challenges and abstract tasks.

Second, we describe the design and implementation of ***Argus***[8], an interactive tool for exploring statistical power, and illustrate how it addresses each of the challenges above. *Argus* is the first direct-manipulation tool that lets researchers (1) dynamically explore the relationships among input parameters such as expected averages or potential confounds, statistical outcome, and power; and (2) evaluate the trade-offs across different experiment design choices.

Third, we describe a **use case** of designing a visualization experiment based on real studies published in TVCG and CHI. The use case illustrates how *Argus* could be used to incorporate information from prior

[8] Argus is openly available at https://zpac-uzh.github.io/argus/

work and explore possible outcome and power scenarios, resulting in an informed decisions for pilot studies and the actual experiment.

Finally, we conducted a **think-aloud study** to assess how *Argus* helps researchers gain insights from exploring relationships among experiment design concepts and statistical power. We found that *Argus* helped both junior and senior researchers to better understand and appreciate the importance of statistical power when conducting controlled experiments.

We view *Argus* as a first step towards an ecology of interactive software tools that improve the rigor of designing and conducting experiments in VIS, HCI, and beyond.

3.11 Acknowledgments

This work is partially is supported by the Innovation Fund Denmark, the BIOPRO2 strategic research center grant № 4105-00020B, the European Research Council (ERC) grants № 695464 “ONE: Unified Principles of Interaction”, and the University of Zurich GRC Travel Grant. We also thank Michel Beaudouin-Lafon for initial feedback and some vision directions in the beginning of the project.

Chapter 4

SPEED: a Flexible Protocol for Planning the Sample Size of HCI Experiments

Controlled HCI experiments are often designed to evaluate the effectiveness of novel interaction techniques or artifacts. However, estimating an appropriate sample size can be challenging due to, for example, a lack of data from prior work. Sequential experimental design (SED) is a method designed to save time, money, and other resources in medical and psychology experiments by stopping data collection early if the effect is stronger or weaker than expected. SED requires pre-determining the effect size and the sample size. In HCI, reliable effect sizes are rare because of the prevalence of small-sample studies and the paucity of replication studies. To determine sample sizes, HCI researchers often rely on heuristics. We introduce SPEED: a novel protocol that incorporates other statistical techniques for systematically estimating the effect and sample size under the mentioned constraints of the HCI literature. We demonstrate SPEED with data from two previously published HCI experiments, provide templates for planning and analysis as R Markdown notebook, and provide a checklist for designing and reviewing SPEED experiments. We also present a web

Publications: The work in this chapter is a collaboration with Wendy E. Mackay, Michel Beaudouin-Lafon, Kasper Hornbæk, and Chat Wacharamanotham. The author is responsible for the use cases, protocol, demonstration, and implementation. This work is currently under revision and is planned for submission to ToCHI. The supplementary materials will be available when the paper is published.

application that eases exploration and comparisons of experimental design candidates. We discuss how SPEED enables researchers to describe and justify nuances and factors influencing their sample-size decisions.

4.1 Introduction

Controlled experiments in the field of Human-Computer Interaction (HCI) usually involve human participants. Determining an appropriate number of participants—or sample size—is an issue that is frequently discussed in the HCI research methods literature, see [Robertson and Kaptein, 2016, Cairns, 2019, Hornbæk, 2013, Lazar et al., 2017a]. Having too few participants leads to statistically *underpowered studies* that cannot detect effects of interest [Rosenthal, 1979]. A non-significant effect could result in not writing or publishing a paper, as papers with null results tend to be rejected. The tendency to reject papers with non-significant results leads to the "publication bias", where published studies make up only a small percentage of the larger number of studies that have been conducted [Rosenthal, 1979]. In the field of HCI, Cockburn et al. [2018] proposed addressing the publication bias by requiring that experimental studies be preregistered. However, no HCI conference or journal has adopted preregistration as a requirement.

A priori power analysis can be used to estimate sample sizes for statistically powerful experiments based on effect sizes. Effect size estimation is challenging in HCI because replication studies are rare [Hornbæk et al., 2014] and undervalued [Greenberg and Buxton, 2008]. Furthermore, sample sizes in HCI experiments, except crowdsourcing studies, usually range from 2-30 participants [Barkhuus and Rode, 2007, Hornbæk and Law, 2007]. Small-sample studies are likely to yield imprecise effect sizes, with a wide confidence interval, or inaccurate effect sizes, that differ from those that actually occur in the population [Gelman and Carlin, 2014]. Shifting from frequentist to Bayesian estimation statistics could allow knowledge to accumulate across studies [Kay et al., 2016b]. However, its adoption has been hindered by the complexity in determining and understanding the choice of prior probabilities as well as the lack of direct replications [Phelan et al., 2019, Sarma and Kay, 2020, Hornbæk et al., 2014].

To help HCI researchers better plan the number of participants, we present SPEED—the Sequential Protocol for Efficient Experiment Design—a comprehensive protocol for choosing the sample size for an experiment. This protocol extends *a priori* power analysis to include sensitivity power analysis and two statistical techniques from the fields of psychology and clinical medicine: Sequential Experimental Design (SED) and Bias- and Uncertainty-Corrected Sample Size (BUCSS). When resource constraints dominate the sample-size decision, the sensitivity analysis can be used to determine how likely a particular study's sample size will detect any effects large enough to be practically useful. When the literature only has small-sample studies, the uncertainty caused by their imprecise effect sizes can be mitigated with the BUCSS extension of *a priori* power analysis. SED helps prevent underpowered and overpowered studies by letting researchers terminate data collection early—without affecting Type I error—, either in cases when the effect sizes are too small to be practically significant or much larger than the initial estimation. We combine sensitivity power analysis, BUCSS, and SED into a principled decision process for sample sizes in controlled experiments (Figure 4.1). This combination of methods specifically addresses effect size availabilities (Table 4.2) and small sample sizes in HCI [Caine, 2016]. These techniques operate within a framework of frequentist statistics, e.g., *t*-test and ANOVA, that many HCI researchers are familiar with. These methods are also compatible with Open Science practices such as preregistration.

Each of these techniques are already widely used outside HCI. With SPEED, we assemble them into a systematic decision protocol suitable for designing HCI experiments. Our goal is to help HCI researchers efficiently maximize their use of limited resources and participant pools. Towards this goal, we present three contributions:

- an introduction to sensitivity power analysis, SED and BUCSS with two detailed examples that illustrate their benefits over traditional approaches for planning HCI experiments;
- R templates and a checklist for authors for reporting studies using SED and BUCSS and for reviewers to assess their rigor; and
- SPEEDX, a web application for researchers explore possible SPEED experimental designs.

4.2 Background and Motivation

Controlled experiments are one of many evaluation methodologies available for HCI research for late-stage designs[1]. Eiselmayer et al. [2019]'s interview study shows that researchers face several challenges when designing controlled experiments, including visualizing and comparing design alternatives. They also find that constraints, such as access to participants, limit the applicability of *a priori* power analysis for sample size planning. According to Wang et al. [2021] analysis, one of the challenges in *a priori* power analysis is the uncertainty in effect size estimation. Indeed, Cockburn et al. [2018] and Cockburn et al. [2020] argue that many HCI experiments are necessarily exploratory because the field has yet to establish a strong empirical foundation and because the technologies and application contexts change rapidly.

Researchers face one of the following cases when planning the sample size for their experiments:

1. The literature reports relevant effect sizes obtained from a sizable amount of samples;
2. The literature only reports effect sizes from small-sample studies; or
3. No applicable effect size is available from the literature.

Only in the first case, researchers can use an *a priori* power analysis to statistically support their decision on the number of participants. In fact, according to our survey of CHI proceedings from 2015 to 2020 in Table 4.1, among the average 420 papers per year that report a controlled experiment, less than ten papers per year (2.3%) use *a priori* power analysis to plan sample size. (See Appendix E for details.)

Even if the field of HCI make a concerted effort to produce systematic reviews and meta-analyses, the resulting effect sizes could still be unreliable. The publication bias cause the effect sizes from these studies to be overestimated [Brand et al., 2008]. Brand et al. [2011]

[1] We agree with Greenberg and Buxton [2008] that controlled experiments are unsuitable for early design iterations [Greenberg and Buxton, 2008, p. 113]. Other legitimate evaluation methods should be used instead [Greenberg and Buxton, 2008, p. 114,117].

CHI full papers	2015	2016	2017	2018	2019	2020	Average
Used controlled experiment(s)	315	369	391	464	474	507	420
Planned sample size with *a priori* PA	1	3	3	8	6	9	5

Table 4.1: Number of papers that use controlled experiments vs. those that used *a priori* power analysis (PA).

found that when the effect sizes are small or medium, the publication bias is the major source of effect-size distortion [Brand et al., 2011]. However, this problem could be mitigated by aggregating the measurements over multiple trials or question items—which are common in HCI, e.g., the average movement time in Fitts's law studies or the workload score from NASA-TLX.

Unlike in other fields of study [Fanelli and Ioannidis, 2013, Carter and McCullough, 2014], the field of HCI has yet to systematically study the prevalence of the publication bias and its effect on the effect sizes. However, there are signs of effect-size inflation: In a meta-analysis of text-entry experiments published at CHI conferences [Obukhova, 2021], standardized effect sizes from 21 research papers were aggregated into small-, medium-, and large-effect groups. The aggregated effect size of the small and medium groups were lower than the corresponding Cohen's d benchmarks [Cohen, 1988, section 2.2.3]; whereas the effect size in the large group exceeded the benchmark. Thus, according to Brand et al. [2011]'s consideration, the text-entry literature is likely to be susceptible to the effect-size inflation—that is large effect sizes show up for small experiments giving a sense that the effects are stronger than they really are. Therefore, even if meta-analyses become widespread in HCI in the future, researchers still need to be vigilant of overestimated effect sizes.

Let us turn to the latter two cases: when only small-sample studies or when no previous studies are available. The uncertainty about or the lack of relevant effect size may cause researchers to underestimate the effect size during an *a priori* power analysis, resulting in sample sizes that are too large to be practical. Thus, some researchers deem *a priori* power analyses irrelevant or only suggestive [Eiselmayer et al., 2019]. Instead, researchers use rules of thumb, e.g., "more than 12" [Eiselmayer et al., 2019, Hwang and Salvendy, 2010] or a local standard within each application domain [Caine, 2016, Hornbæk and Law, 2007].

Running an experiment with too few participants is likely to lead to non-significant effects, which consequently contributes to the publication bias and may encourage HARKing [Cockburn et al., 2018, 2020] (Hypothesizing After the Results are Known). Running an experiment with too many participants, on the other hand, can be seen as a waste of resources: participants, time, and money.

This situation, however, is constrained by an *assumption that the sample size must be set before collecting any data*. But this fixed-sample assumption is, in fact, not necessary, as demonstrated by techniques that have been in use in the fields of medicine and psychology for a long time already.

4.2.1 Techniques for Adapting the Design of On-going Experiments

According to textbooks on experimental design in HCI—such as [Lazar et al., 2017a, p. 459] or [Purchase, 2012, p. 78]—sample size decisions must be finalized before data collection. This practice of *fixed-sample design* (FSD)[2] is designed to control the probability of committing a Type I error (α) with statistical tests that analyze the whole dataset Neyman [1942, 1956], Pearson [1955]. Other methods for choosing the sample size include sequential experimental design (SED), adaptive trial design (AD), and Bayesian experimental design (BED). We describe each of them below and summarize the flexibility each method provides in Table 4.2.

	FSD	SED	AD	BED
Stopping before the full sample size	✗	✓	✓	✓
Increasing the sample size beyond the initial plan	✗	✗	✓	✓
Changing experimental conditions to during the experiment	✗	✗	✓	✓
Running multiple statistical analyses before the end of the experiment	✗	✓	✓	✓
Simple additions to preregistration compared to FSD	–	✓	✗	✗
Requires data monitoring committee for transparency	✗	✗	✓	✓

Table 4.2: Flexibility and practicality of experimental design methods.

[2]The term fixed-sample design was first used by Pocock [1977] in his paper that proposes sequential experimental design. This term is still used in contemporary literature such as the FDA guideline [Food and Administration, 2019].

Sequential Experimental Design (SED)

Instead of conducting one statistical test after finishing data collection, researchers can run the test at multiple intervals as data accumulate, while controlling for the probability of Type I error. In 1969, Armitage et al. [1969] showed that this control could be achieved by lowering the α level for each of the tests. Subsequently, in 1977, Pocock [1977] proposed the sequential experimental design procedure, which lets researchers plan to analyze data at pre-specified intervals and stop an experiment early when the effect sizes are much smaller or much larger than anticipated. SED can be applied to design studies that use any counterbalancing design and is considered a routine practice in medical experiments [Gaydos et al., 2009]. Pocock [1977]'s procedure and its subsequent improvements have been called "group-sequential design", "sequential analysis", or "group sequential trial". We use the term "sequential experimental design" to avoid confusion with terms that HCI adopts from psychology.

Adaptive Trial Design (AD)

Adaptive trial design methods, including the seminal work by Bauer [1989], enable researchers to modify the design of an on-going experiment by incorporating information that becomes available after the experiment starts—e.g., from data collected so far or data from external sources. Adaptive trial designs enable researchers to add or drop independent variables or their levels, change the magnitude or dosage, change the number of trials, or change counterbalancing strategies [Pallmann et al., 2018]. AD and SED were developed separately until it was recognized that SED is a special case of adaptive trial designs pBauer et al. [2016], Food and Administration [2019]. Researchers can also add AD to an existing SED study [Müller and Schäfer, 2001]. To ensure that researchers make appropriate choices, AD studies require researchers to work closely with a data monitoring committee throughout each experiment [Gaydos et al., 2009]. Reporting the results of AD studies is also challenging because there is no method for calculating point and interval estimations (e.g., mean and confidence intervals) that is generalizable to all AD designs [Pallmann et al., 2018].

Bayesian Experimental Design (BED)

The use of the word "Bayesian" in experimental research can be ambiguous [Campbell, 2013] as it can mean Bayesian analysis or Bayesian experimental design. *Bayesian analysis* uses information that is available before the experiment—e.g., from previous experiments in the literature—to create a set of prior probability distributions to be used in statistical models during data analysis. This method can be applied to fixed-sample, sequential, or adaptive experimental designs. For fixed-sample designs, several works have introduced Bayesian analysis to HCI, e.g., [Kay et al., 2016b, Phelan et al., 2019]. Bayesian analysis can also be used with SED experiments: Stopping rules can be defined in a Bayesian manner [Schönbrodt et al., 2017], and at the end of the experiments, researchers can calculate credible intervals or Bayes factors. The resulting credible intervals or Bayes factors do not require additional adjustments [Jennison, 1999, section 18.1]—unlike frequentist SED (see Section 4.4.6).

By contrast, *Bayesian experimental design* is a set of decision-making procedures for AD experiments [Lai et al., 2012, Food and Administration, 2019]. It requires experimenters to define a utility function that guides design choices before starting to collect data. While an experiment is on-going, this method uses the information collected so far to optimize design choices for the remainder of the experiment [Chaloner and Verdinelli, 1995]. Experiments that are designed with this method can use Bayesian or frequentist statistical analysis. BED does not guarantee the probability of Type I and II error, so researchers must estimate them with a simulation. The simulation procedures are currently only available for parametric models—e.g., requiring the normality assumption [Lai et al., 2012].

4.2.2 Choosing a Suitable Method for Designing HCI Experiments

SED, AD, and BED each enable the sample size to be flexible, a property that we have shown to be desirable for HCI experiments. AD and BED also enable the flexibility of adjusting experimental conditions, but at a higher cost and complexity in planning, monitoring, and analysing the experiment. We argue that the flexibility in experimental conditions is rarely necessary because HCI research can narrow them down by using, e.g., preliminary studies and pilot studies.

These pre-experiment studies can also collect richer data, e.g., with qualitative methods, to help refine the experimental design further.

SED is simpler than AD and BED, making it easier for reviewers and readers to assess the validity and transparency of design decisions. It addresses the typical HCI problem of effect size uncertainty in small-sample studies, and it can help in situations where the sample size is the primary limitation in the design, e.g., when requiring a population with specific characteristics.

[Lakens, 2014a] argued for adopting SED for controlled experiments in psychology. His article offers a tutorial for basic two-condition between-subjects design, a guide on using a GUI application for planning, and an Excel spreadsheet to compute adjustments for the statistical results. Our work further facilitates the adoption of SED in three ways:

1. We provide templates that cover within-subjects or mixed experimental designs.
2. Our templates are reproducible and can be preregistered.
3. We characterize the choices of spending function and discuss how to choose one.

In a broader view, our paper also situates SED in a broader experimental design process to address the characteristics of the CHI literature as described in Section 4.2.

In the next section, we propose how SED and two other techniques can be used systematically to make sample size decisions more flexible. We describe the additional information needed by SED, provide recommendations for authors and reviewers, and offer a checklist for using the process we propose.

4.3 Motivating Use Cases

Our protocol extends the researcher's toolbox with a tool that can save participants when they are not absolutely necessary. Saving participants is particularly useful in the following three use cases: large on-

line studies, studies with hard to access participants, and studies replicating small-sample studies. We present these three use cases along with the challenges that researchers face in these situations. Section 4.5 describes how SPEED addresses them.

4.3.1 Large Online Studies

Online studies tend to be easier to scale up than studies conducted in the lab. However, researchers often need to recruit extra participants to address the possibility that data from some participants needs to be dropped, e.g., if participants fail attention-check tasks.

For example, Hofman et al. [2020] wanted to understand how participants estimate the magnitude of an effect when representing it with either a confidence interval or a prediction interval. The study included attention-check tasks to exclude data from participants who failed them. Based on a pilot study, the researchers conducted an *a priori* power analysis yielding a sample size of 1,700 participants. The pilot study also revealed that roughly 30% of participants failed the attention-check tasks. Thus, the researchers estimated the upper limit for the sample size to be N = 2,400 and proceeded to recruit that many participants. However, it is difficult to figure out how large the sample size actually needed to be.

Challenge: To plan the sample size of an online study, researchers need to consider the possibility that the sample size yielded by the power analysis needs to be increased to accommodate poor data quality. With SPEED, researchers can plan online studies with an increased sample size, but data collection can be stopped early if the data quality is better than expected.

4.3.2 Studies With Hard to Access Participants

Lab studies are a common evaluation method to assess artifacts in HCI. Judicious use of participants is desirable in the following circumstances:

1. when the participant pool is limited, e.g., in accessibility studies;

2. when participants' involvement in a study renders them ineligible to be included in future studies, e.g., studies that involves deception may make the participants suspicious of subsequent similar studies [MCGRATH, 1995, Hornbæk, 2013]; and

3. when access to participants is difficult because they are very busy, e.g., surgeons [Avellino et al., 2021].

In these circumstances, researchers have to weigh these downsides with the benefits of having a larger sample size, i.e. higher statistical power.

For example, Wu et al. [2014] wanted to understand if their interactive checklist system helped medical staff to be more effective in responding to crises. They planned a controlled experiment to compare their system with the existing paper-based checklist and no checklist at all. It was difficult to estimate how many medical staff should be recruited for the evaluation, and the experimenters were aware that the time of the medical staff was precious. In the end, having spent two years designing their system, they decided to recruit all available medical doctors and practicing medical students from their participant pool (N = 37) in order to maximize the chances of getting a significant result.

Challenge: To plan the sample size of a study where judicious use of participants is important, researchers need to trade off the sample size with the desired confidence in the statistical findings. With SPEED, researchers can plan the study with the whole participant pool, but stop data collection if the effect is much stronger than expected.

4.3.3 Replicating Small-sample Studies

Replicating and extending existing studies is common in experimental fields, but is not as widespread in HCI even though it could strengthen the validity of previous results and help improve artifacts and interaction techniques. However, when replicating a small-sample study, researchers should consider recruiting a larger sample than the original study.

For example, earlier work has explored how users can enter sensitive information in public settings. Khamis et al. [2018] compared

the performance of 17 users entering a PIN with three input techniques: touch, mid-air gestures, and eye gaze, and found that touch was fastest. Mathis et al. [2021] wanted to find out if the same results obtain in a virtual reality (VR) setting, so as to make it easier to run future similar studies. It was difficult to figure out the number of participants to use in a replication of Khamis et al. [2018] because in addition to the normal considerations about sample size, the estimate of the effect size might be inflated due to the low number of participants in the original study. In the end, Mathis et al. [2021] proceeded to replicate with the same (low) number of participants as the original study (N = 15).

Challenge: To plan the sample size of a replication study based on previous small-sample studies, researchers need to consider the possibility that the original effect size is overestimated. With SPEED, researchers can use a power analysis that adjusts for small-sample studies, and stop running participants if the effect replicates as expected.

4.4 Sequential Experimental Design and Sample Size Adjustment with SPEED

Sequential experimental design and sample size adjustment are two procedures that expand fixed-sample design. To introduce these concepts, we first give an overview of SPEED, and then explain these two procedures in detail.

4.4.1 Overview

According to the interview study by Eiselmayer et al. [2019] about fixed-sample experimental design, researchers start with the conceptualization of the design and specification of hypotheses. Next, they explore possible counterbalancing designs, which determine the number of replications, the blocking, and the counterbalancing strategies. While iterating the experimental design, researchers sometimes conduct an *a priori* power analysis to inform the sample size, i.e. the number of participants. After the study is planned, they proceed to collect all data, conduct the analysis, and report its results (Figure 4.1, left).

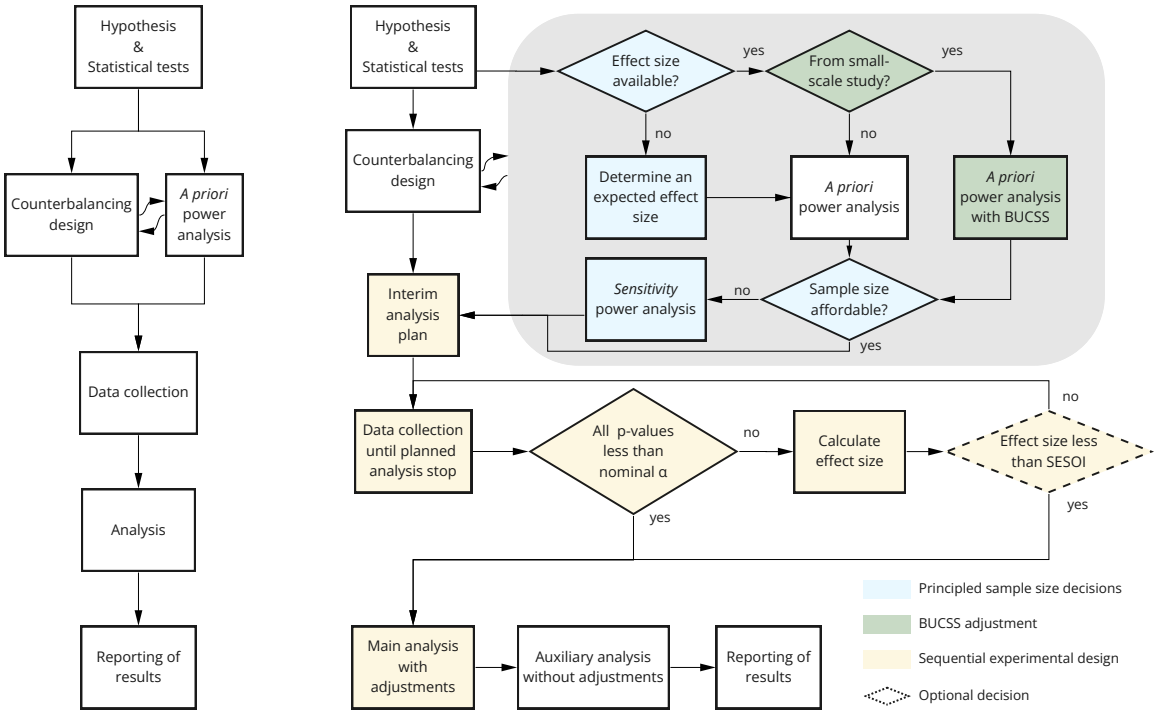


Figure 4.1: The process diagram of SPEED with SED and sample size adjustment. The dashed decisions are optional.

SPEED extends the sample size planning, data collection, and analysis. Figure 4.1 (right) shows an overview of the whole process with unchanged components in white, sample size adjustment components in green, and sequential experimental design (SED) components in yellow. Because of their synergies, we present them both as an integrated process, but each could be used independently from each other. As with fixed-sample design, researchers start with the conceptualization of the design. Next to the counterbalancing design, which remains identical, researchers conduct one of three different power analyses to determine the sample size or the smallest effect size of interest (SESOI) for the study (Section 4.4.2). The choice depends on the availability of related effect sizes. The counterbalancing design, the sample size, and the SESOI are used to plan the interim analyses with the respective boundaries for halting or commencing the monitored data collection (Section 4.4.3). Researchers collect data until an interim sample size is reached. At this point, researchers perform the statistical analysis for the main hypothesis (Section 4.4.4). If all results are statistically significant, researchers can stop the data collection. If at least one result is not significant, researchers can decide whether or not to continue

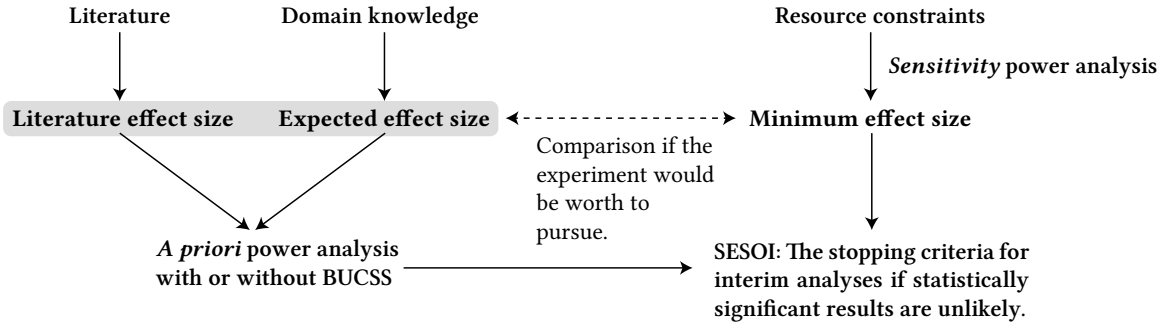


Figure 4.2: Three different terms for effect size are used during sequential experimental design. The origin and their usage are outlined.

the study based on the effect size. After stopped data collection, researchers adjust the analysis results and conduct auxiliary analyses such as an assessment of order or interaction effects (Section 4.4.6).

4.4.2 Power Analysis

In fixed-sample design, researchers conduct an *a priori* power analysis with an effect size and a statistical power to determine a single target sample size. The SPEED protocol incorporate several common situations in HCI: when the effect sizes are unavailable in the literature or are based on small-sample studies, or when the resource constraint is the hard limit. The choice of the procedure depends on the type of effect size available as described below:

Types of Effect Sizes

Five types of effect sizes are relevant to the SPEED protocol. The overview of their relationship is shown in Figure 4.2.

A **reliable effect size** is an effect size from a similar study that can be taken to plan the new study. Such study should have a relatively large sample size—at least according to the standard of the sub-field, e.g., [Abbott et al., 2019, Caine, 2016]. Alternatively, a reliable effect size can be calculated by aggregating the results of multiple studies with a meta-analysis, e.g., [Obukhova, 2021]. This effect size is used as input for the *a priori* power analysis (Section 4.4.2).

An **unreliable effect size** is also from a similar study, conducted with small sample sizes. Online studies aside, HCI experiments use relatively small sample sizes [Barkhuus and Rode, 2007, Hornbæk and Law, 2007]. Effect sizes in small-sample studies are likely to be overestimated and sometimes even estimated in the wrong direction [Gelman and Carlin, 2014, Maxwell, 2004]. Overestimated effect sizes can lead to significant findings even though the effect might not exist in the population. The overestimation can be aggravated by the publication bias (Section 4.8.3). With an unreliable effect size, the SPEED protocol suggests using BUCSS (Section 4.4.2).

An **practical effect size** stems from the researcher's domain knowledge and is estimated by assessing the practical significance. An effect size has a *practical significance* when it is large enough for the finding to be useful [Kirk, 1996]. Researchers are encouraged to subjectively determine the level of practical significance and support this decision with a sound argument—for example, by weighing the effect with the cost and scalability of the tested intervention [Kirk, 1996, Dragicevic, 2016, Bakker et al., 2019]. Researchers may substitute the reliable effect size with a practical effect size for the *a priori* power analysis (Section 4.4.2). The practical effect size is also relevant in planning under resource constraints in sensitivity power analyses (Section 4.4.2).

When the resource or participant constraints is a dominant concern, instead of planning for a certain level of statistical power, researcher may aim to determine the **minimum effect size** that the experiment could likely detect. This minimum effect size is compared to the other types of effect size above to determine whether running the experiment will likely be futile or not. This procedure is explained in sensitivity power analysis (Section 4.4.2).

The **smallest effect size of interest (SESOI)** is used in sequential experimental design (Section 4.4.3). It is as threshold where researchers may decide to stop the data collection before reaching the full sample size when the study is unlikely to yield any statistically significant results. The choice of SESOI depends on the type of power analysis, as described below.

A priori Power Analysis

If there is an effect size from a previous study that use a relatively large sample size (or from a meta-analysis, researchers can directly conduct the *a priori* power analysis. Here, researchers choose a threshold of statistical power (usually .8) and find the sample size that will likely to yield the power exceed that threshold.

If the literature effect size is unavailable, researchers can still conduct *a priori* power analysis with an expected effect size. In this case, researchers should provide sufficient argument to support this decision in their paper. It is advisable to err on the low side because studies that can detect a small effect size will be able to detect the larger ones [Murphy et al., 2014, p. 21].

A priori Power Analysis with BUCSS

As described in Section 4.4.2, effect sizes from small-sample studies risk overestimating the population effect size, and studies that did not achieve statistical significance are left unpublished due to publication bias. Taylor and Muller [1996] created a mathematical model of both problems, and Anderson et al. [2017] implemented this model in the R package `BUCSS` [Anderson and Kelley, 2019]. We describe the intuition behind this model below.

Consider a long sequence of replication studies. The sample size of each study is planned based on the results of the previous one. When a study yields a larger-than-expected effect size, the subsequent study will use a smaller sample size to reduce the excess statistical power. Eventually, smaller sample sizes will lead to an underpowered study, which is left unpublished due to publication bias. Thus, the result of this underpowered study will not be used to upward-correct the sample size of the following study. Eventually, one of the later studies will yield a large effect size by chance, allowing the next study's sample size to be increased, and the process repeats. Taylor and Muller [1996]'s model encodes the censoring due to publication bias and the upward-downward corrections in two parameters: α_P and assurance:

$\boldsymbol{\alpha_P}$ represents how much study results are censored by publication bias. $\alpha_P = 0.05$ represents the situation that all studies that yield

p-value higher than 0.05 are left unpublished. $\alpha_P = 1$ means no publication bias.

Assurance represents the proportion of these hypothetical studies that overestimate the population effect size. The assurance level of 0.5 represents a balance: half of the replications overestimate the effect size, which is corrected by the other half, resulting in underestimations. Higher assurance levels protect against overestimations.

What level of assurance should be used? Taylor and Muller [1996] suggest a conservative assurance level of 0.95—which usually result in a very high sample size. Anderson et al. [2017] conducted a simulation study with unpaired t-tests, paired t-test, 3×2 between-subjects ANOVA, and 3×4 mixed-model ANOVA. These simulated studies are planned based on the original study that uses a sample size of 25. The results indicate that setting assurance to 0.8 is sufficient for the medium population effect size according to Cohen's criteria. In practice, the population effect size is unknown. The available information is the unreliable effect size and the sample size of the prior study.

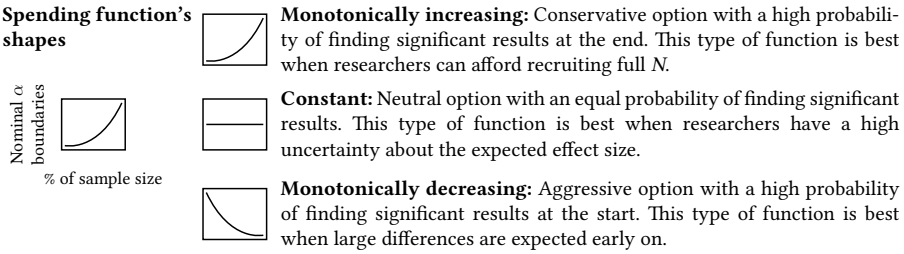
Recommendation: Start with the assurance of 0.8. When BUCSS suggests an affordable sample size and if the prior study has a very low sample size, extraordinarily large effect size, or both, consider increasing assurance liberally up to 0.95. When BUCSS suggests a sample size that exceeds the available resources, we recommend lowering the assurance cautiously because reducing assurance has a diminishing effect on sample-size reduction. The lowest assurance level is 0.5—which does not correct for any overestimation. For α_P, we believe that 0.05 is reasonable because the field of HCI neither enforces preregistration nor registered reports.

Sensitivity Power Analysis

When the sample size suggested by the *a priori* power analysis exceeds the available resources, researcher can use the largest sample size they can afford to determine the smallest effect size that they can detect. This method is called *sensitivity* power analysis (Cohen [1988], p. 15; Murphy et al. [2014], pp. 86–87; Lakens [2022], p. 14). Researchers specify the largest affordable sample size and statistical power to obtain the minimum effect size that is likely to be detectable. Then, researchers compare this minimum effect size to the benchmark effect

size (either unreliable, reliable, or practical effect size). When the minimum effect size is far smaller, researchers should reflect whether to pursue the experiment with a risk of futility. If the minimum effect size is acceptable, it could be used as the SESOI for the next step.

4.4.3 SED Plan



Spending function	Shape & parameter	Notes
Pocock 1977		• Requires equal spacing • Supports only one-sided statistical tests • Not parameterized • Not used anymore
O'Brien- Fleming 1979		
Lan-DeMets 1983 "α spending function"	ρ [0, 0.35] (0.35, 2]	Supports only one-sided statistical tests
Kim-DeMets 1987 "power family"	ρ [0, 1) 1 >1	Extension of Lan-DeMets to support symmetric and asymmetric two-sided boundaries
Hwang-Shih-DeCani 1990	γ >0 $=0$ <0	Generalization of Lan-DeMets and Kim-DeMets
Exponential 1987	ν [0, 0.35] (0.35, 1.5]	One- or two-parameter function with greater flexibility

Table 4.3: Common spending functions for the probability of committing Type I error (α) [Pocock, 1977, O'Brien and Fleming, 1979, Lan and DeMets, 1983, Kim and DeMets, 1987, Hwang et al., 1990, Anderson and Clark, 2009].

The SED plan includes information about the main hypothesis, statistical procedures, interim analyses, and stopping criteria. Next to the main hypothesis, there might also be hypotheses that are not answering the main research question which can be excluded from the SED plan. For example, a nuisance variable such as the handedness of par-

ticipants is included to ensure that it does not confound the results. Researchers can set up the interim analyses with information about the counterbalancing of the experiment, the sample size, the main hypothesis, and statistical procedure.

Researchers decide the sample sizes at which interim analyses will be conducted, i.e. information time τ. For example, with a total of maximum 60 participants, suppose we decide to conduct four analyses at information times $\tau = \frac{1}{4}, \frac{2}{4}, \frac{3}{4}$, and $\frac{4}{4}$. Three interim analyses would be performed at 15, 30, and 45 participants. The final analysis would then be performed at 60 participants. Positioning the analyses must take the counterbalancing into account, e.g., interim analyses for a 2×2 Latin square counterbalancing design need to occur at a sample size that is a multiple of 4. Box 1 summarizes considerations when choosing the number of analyses.

Box 1: Considerations in choosing the number of analyses

There are several considerations to take into account when choosing the number of analyses. First, it is crucial to perform an (interim) analysis only at the point where the number of participants reaches a fully counterbalanced unit. Second, conducting a greater number of analyses reduces the nominal α threshold for each analysis, consequently decreasing the overall likelihood of obtaining statistically significant results. Third, conducting more analyses increases the potential cost savings associated at each analysis.
It is important to note that there is no definitive or correct number of analyses; the optimal choice depends on the specific experimental setting. To give a general idea about the number of analyses, Todd et al. [2001] recommend between 4 and 8 interim analyses for practical considerations. Nonetheless, we encourage researchers to explore different number of analyses and sample size intervals with their respective nominal α thresholds, as outlined in Supplementary Materials Section 3 (S3).

In Null Hypothesis Significance Testing, an α value of 0.05 is commonly used to establish the significance of an effect. While some researchers have proposed to lower the α value to 0.005 Benjamin et al. [2018], others have opposed it [Lakens et al., 2018]. However, researchers can freely choose which ever α value is appropriate for their field as SED maintains the overall α value constant across all analyses. Due to the multiple analyses which are conducted throughout the

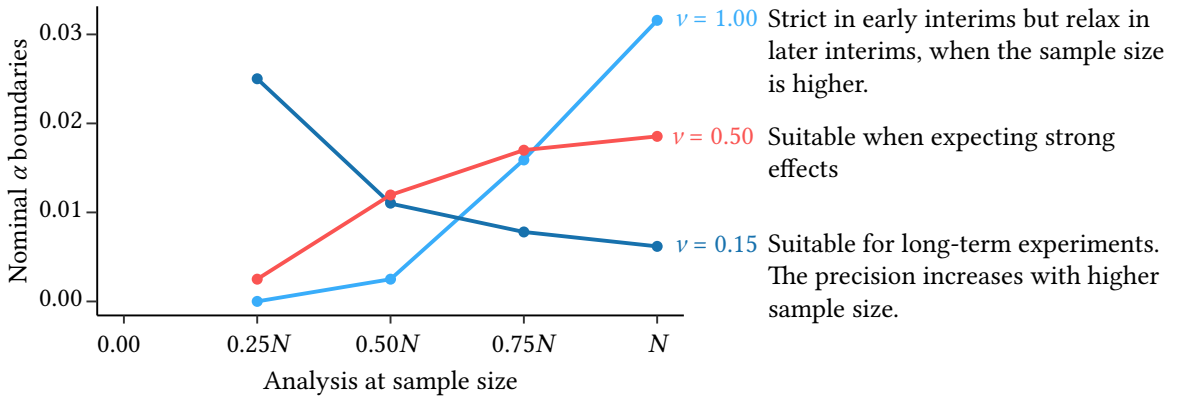


Figure 4.3: Anderson and Clark [2009]'s exponential spending function at three different input parameter settings with four analyses outputting the nominal α boundaries as significance threshold at each analysis.

data collection, the overall α value is distributed to each of the analyses with a spending function. Such a spending function distributes α values over the planned analysis so that the overall Type I error rate is kept at .05. Before Lan and DeMets [1983] introduced the concept of spending functions in 1983, researchers used the procedures by Pocock [1977] or O'Brien and Fleming [1979] to distribute the α value. These two procedures require an equal spacing between the analyses, making them less versatile in experiment settings. Today, researchers can choose from many different parameterized spending functions. Table 4.3 shows common spending functions and their characteristics. See Box 2 for our recommendation on choosing a spending function.

Table 4.3 shows different spending functions and how their parameter setting affects the three possible types: constant, monotonically increasing, and monotonically decreasing. A constant spending function is advantageous if the outcome uncertainty is high as it distributes the α boundaries equally across each interim analysis. A monotonically decreasing spending function is advantageous if the p-value is expected to be high during early stages of the experiment (e.g., Figure 4.3 $\nu = 0.15$). A monotonically increasing spending function is advantageous if researchers expect significant results in the later stages of the experiment, yielding a higher probability to conclude with significant results compared to the other functions (e.g., Figure 4.3 $\nu = 1.00$).

After choosing a spending function, a SED plan looks like Table 4.4, which uses an exponential spending function [Anderson and Clark,

2009]. This table can be preregistered to establish the rigor of the study for reviewers and authors (see Section 4.6 for more details).

	Sample Size at Information Time τ	z-Score	Nominal α
Interim Analysis #1	0.25 N	± 4.80	7.8125×10^{-7}
Interim Analysis #2	0.50 N	± 3.02	1.2492×10^{-3}
Interim Analysis #3	0.75 N	± 2.22	0.0134
Final Analysis	1.00 N	± 1.81	0.0358

Table 4.4: SED plan using Anderson and Clark [2009]'s exponential spending function with 80% power and $\nu = 1.00$.

Table 4.4 includes a z-score—Also known as standard score.—next to the nominal α value. The z-score represents the distance in standard deviations of a measurement to the sample mean and is defined as:

$$z = \frac{x - \mu}{\sigma}, \tag{4.1}$$

where x is the measurement, μ is the sample mean, and σ the sample standard deviation. Test statistics such as t, F, or χ^2 can be converted into p-values which in return can be converted into z-scores. Computing the z-score directly has the benefit of including the direction of the effect, i.e. whether it is positive or negative. As a p-value lacks this information, researchers can add the sign (+ or -) onto the z-score calculated from test statistics. For planning SED studies, it is beneficial but not necessary to use the z-score along with α and p-values as the sign is preserved and the readability increased for small p-values.

Box 2: Choosing Sample Size Intervals and a Spending Function

When choosing a certain spending function, researchers need to trade off the benefit of stopping early against the risk of not finding significant results at the final sample size. At later stages of the experiment with a larger sample size, the precision improves and it is more probable to get lower p-values provided the effect exists in the population. Choosing a monotonically decreasing spending function (Table 4.3) has the benefit of stopping early with a comparatively high probability. However, the risk that early p-values might be large due to inter-participants variability, hence, preventing early stopping makes it unlikely to find significant results later. Therefore, we do not recommend monotonically decreasing spending functions, and instead focus on constant and monotonically increasing functions. Constant functions are advantageous if researchers have no related work that could provide an idea about what the outcome might look like. Monotonically increasing functions are advantageous if researchers want to focus on finding significant results in later stages of the experiment. Without additional information about the experiment outcome, we recommend using Anderson and Clark [2009]'s exponential spending function with $\nu = 1$ for conservative experiments (Figure 4.3 light blue plot).

4.4.4 Data Collection and Interim Analyses

After the SED plan is finished, researchers can start the data collection that will have three possible outcomes: an early stop due to significant results, an early stop due to small effect sizes, and a regular stop. When the first interim analysis sample size is reached, researchers conduct the analysis on the primary hypotheses. If the p-values for each tested comparison is below the nominal α boundaries of the interim analysis plan, researchers can stop the data collection and proceed with significant results. This early stop indicates that the main effect of interest is much stronger than anticipated and researchers can save further resources.

If at least one of the p-values is larger than the nominal α boundary, i.e. not statistically significant, it is possible that the real effect size

is too small to be detected, and further data collection would be futile. To determine this situation, researchers proceed by calculating the mean effect size (ES) and compare it to the SESOI. If the ES is much smaller than the SESOI, researchers can choose to stop the data collection. This early stop indicates that the main effect of interest is much smaller than anticipated and researchers will unlikely find significant results even with the full sample size. In HCI, this could imply that a novel interaction technique is not as good as predicted or that the experiment is not set up well enough to capture its qualities.

If at least one p-value is larger than the nominal α boundary and the ES is larger than the SESOI, researchers can proceed with the data collection until the next interim analysis or until the final sample size is reached. In the latter case, the experiment comes to a normal stop, which is the case for experiments where the initial effect size was a good estimate of the sample effect size.

SED does not prevent researchers from HARKing or reporting a planned SED as a normal study. However, if researchers throw away the SED plan and report the study as is, they risk obtaining an imprecise effect size, i.e. a wider 95% confidence interval, calling the results into doubt. SED is compatible with preregistration which addresses publication bias and HARKing [Cockburn et al., 2018]. We listed the required information for preregistering a SED study in point 6 of the checklist in Section 4.6.

4.4.5 Multiple Comparison Adjustment

If multiple hypotheses are tested at each interval, researchers should adjust each nominal alpha to account for multiple comparisons, e.g., with the Bonferroni correction ($\frac{p}{m}$; m is the number of hypotheses) [Bretz et al., 2009, Jennison, 1999]. However, the Bonferroni correction might be too conservative, especially because the nominal α is already small. Instead of reducing α for all hypotheses to the same amount, stepwise procedures such as the Benjamini-Hochberg correction [Benjamini and Hochberg, 1995] would be more practical.

4.4.6 Data Analysis and Result Adjustments

Once data collection is completed, the observed effect sizes and p-values could be biased [Lakens, 2014a, Jennison, 1999]. Researchers can choose to adjust p-values, mean differences, and confidence intervals. Researchers can choose to adjust[3] p-values, mean differences, and confidence intervals. However, in the SED literature, there is no consensus whether this adjustment is mandatory and a vital part of SED. On the one hand, Dupont [1983] argues that the unadjusted p-values[4] should be reported as it represents the strength of the evidence based on the underlying data. Pocock [2005] argues for reporting the unadjusted p-value for the sake of simplicity. On the other hand, Proschan et al. [2006] argue for the adjustment because of the potential overestimation in the data because it was analyzed while being collected. Below, we summarized the rationale for these adjustments and provide a recommendation for reporting adjusted parameters in Box 3. Mathematical details are provided in Appendix F.

Adjusting p-values

In fixed-sample design, a p-value is *"the probability of obtaining an effect that is at least as extreme as the observed effect assuming that the null-hypothesis is true"* Proschan et al. [2006]. This definition needs to be adapted to fit the monitored data collection during SED: The adjusted p-value is *"the probability of obtaining an effect that is at least as extreme as the observed effect assuming that the null-hypothesis is true* ***while not observing a significant difference at earlier interim analyses.***" Lakens [2014a]. Hence, an adjusted p-value is the union of probabilities that the z-score does not exceed the nominal α boundary at earlier interim analyses and the probability of exceeding the boundary at the final analysis.

[3] The adjustment discussed in this section is in the SED context. If multiple hypotheses were tested, additional adjustments for multiple tests need to be done separately.

[4] Also called observed p-values or nominal p-values in the literature.

Adjusting the point and interval estimations

SED is mathematically modelled as a Brownian motion stochastic process, which is parameterized by a drift parameter. The adjustment of point estimates, e.g., mean differences, and interval estimates, e.g., confidence intervals, take this drift parameter into account.

Box 3: Reporting of adjusted parameters

As there exists no consensus among statisticians whether all results should be adjusted or not, we recommend the following. Report both, the adjusted and unadjusted, p-values. This will facilitate the comparison with the conventional $\alpha = .05$ [Lakens, 2014a]. Adjusting the mean difference and its confidence interval can make the results more conservative, i.e. the mean difference will be closer to 0 and the confidence interval will be wider [Proschan et al., 2006]. As there might be side-effects, e.g., the mean difference can be slightly outside the confidence interval, during the adjustment, we recommend that researchers always report the unadjusted mean and unadjusted confidence interval. The adjusted mean and adjusted confidence interval should also be reported with a caveat to readers that they are more conservative. Nonetheless, we encourage researchers to explore the latter adjustment results and include them as supplementary materials. In general, the results should follow statistical reporting practices that include not only p-values but also confidence intervals and effect sizes [Lakens, 2019].

Auxiliary analyses

Once data collection is finished, researchers can conduct auxiliary analyses that provide further insights into the data. Because these analyses take place only once at the end of the data collection, their probability of Type I error remains the same as in fixed-sample designs. Their results do not need to be adjusted. Further exploratory analyses can also be run without additional restrictions.

4.4.7 R Template for Power Analysis and Sequential Experimental Design

To the best of our knowledge, there is no comprehensive tool or code template available covering the entire process to support researchers in conducting SPEED experiments. However, there are some R packages available that allow researchers to plan and conduct SPEED experiments. Nonetheless, Table 4.5 shows challenges we identified that might prevent HCI researchers from conducting SPEED studies correctly using the respective packages.

Package	Challenge
`pwr` a priori & sensitivity PA	Specification and interpretation of the sample size parameter for within- and mixed-participant designs.
BUCSS a priori PA for small-n	(1) Specification of t.observed if non-t-statistic is reported. (2) Specification of assurance parameter is ambiguous.
`gsDesign` SED planning	Missing result adjustment capabilities.
`GroupSeq` SED planning & adjustment	(1) GUI-based application (2) Missing code documentation

Table 4.5: Challenges for non-statisticians using relevant packages for SED.

Lakens [2014a] created a guide for Psychologists on how to plan a SED study and adjust its results using `pwr`, the GUI of `GroupSeq`, and Microsoft Excel. His guide features multiple scenarios focusing on between-subject designs. However, many studies in HCI are conducted using within- or mixed-participants designs, thus, making the guide less applicable to HCI. Additionally, researchers following Lakens [2014a]'s guide rely on GUI parameter specification and output, making the results less replicable and explorable. Therefore, we argue that current tools and guides do not adequately support planning, conducting, and analyzing SPEED studies according to the HCI-tailored process we propose in Section 4.4.

Phelan et al. [2019] created a set of R templates for conducting Bayesian analysis which were evaluated during a user study. Inspired by Phelan et al. [2019], we used the following three design goals when creating our R template in markdown format that enables researchers to follow our process:

DG1: Allow users to easily plan and conduct a SPEED study with sample size planning with no prior knowledge.

DG2: Communicate statistical parameter specification and interpretation of functions in use. This is particularly important for a HCI audience, which comprises people with a range of expertise in statistics.

DG3: Prioritize the particular needs of HCI researchers. The template should support the analyses that are most relevant to HCI researchers.

In the template, we briefly explain the concepts of power analysis, SED, and result adjustment (DG1). For the functions, we describe how they should be used (DG1) and how to specify certain parameters (DG2). The parameters required to change are marked with "CHANGE ME" and an id linking them to their respective explanation. We interpret the results of the function output and point out what implications follow (DG2), for instance, stopping data collection early. The sample code in the template features a within-participants design that can be extended to a mixed-design (GD3). The simulation study of Hofman et al. [2020] in Appendix G features a between-participants study for completeness. Thus, we believe that our template provides a vital starting basis for HCI researchers who aim to conduct a SPEED study.

Figure 4.4 shows the structure of the R Markdown Notebook template, along with the information researchers need to provide, the used R packages, and our contributions.

Sample Size Planning

We use `pwr` [Champely, 2018] for *a priori* and sensitivity power analysis. We contacted the authors of the package to clarify the ambiguity about specifying the input parameter `n` (the sample size) when planning with a within-participant factor. Interpreting this parameter incorrectly would lead to a sample size that is significantly higher or lower than necessary. For `BUCSS` [Anderson and Kelley, 2019], users need to specify a value for the assurance. As this specification introduces more uncertainty for the users, we provided a guide for researchers on how to specify assurance in Section 4.4.2 and the template.

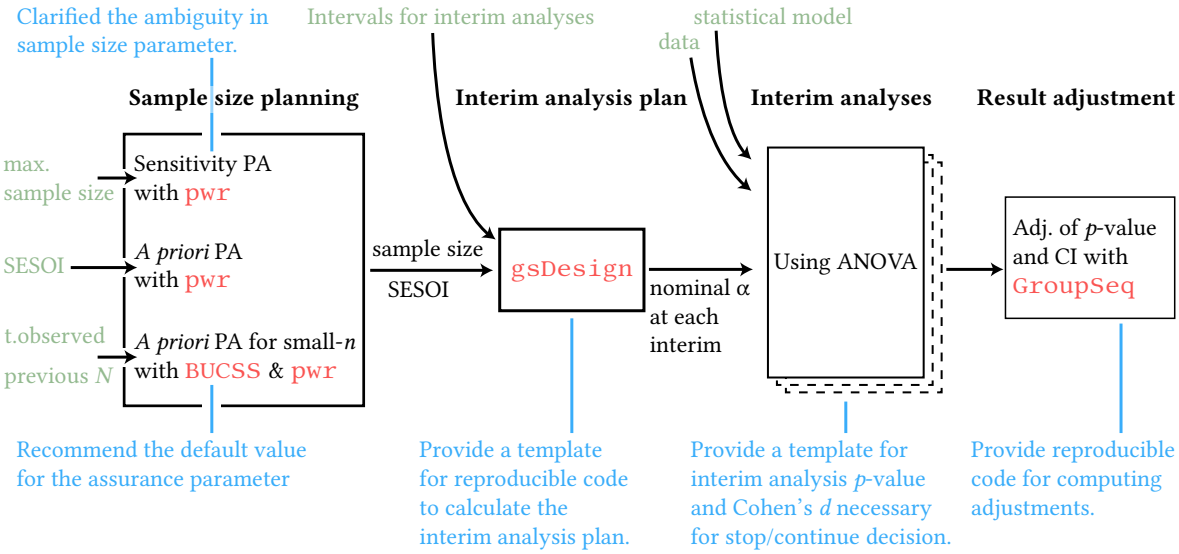


Figure 4.4: The structure of the code template with the four main parts. Inputs by the researchers are highlighted in green, R packages in red, and our contribution in blue. The *p*-value adjustment uses formulas obtain from [Proschan et al., 2006].

Interim Analysis Plan

`GroupSeq` [Pahl, 2018] and `gsDesign` [Anderson, 2020] can both be used to plan a SED study. `GroupSeq` depends on a GUI, so researchers have to take screenshots in order to explore parameter settings, preregister their study, and add the plan to the supplementary materials. `gsDesign` includes a wider set of spending functions and more complex adaptive trial designs. For example, `gsDesign` support planning SED studies with one-sided tests or uneven tails. Additionally, `gsDesign` is well documented—something that is missing for `GroupSeq`—which enables researchers to move to more complex design more easily.

Result Adjustment

`gsDesign` does not include any functions to perform the result adjustment. `GroupSeq` supports the calculation of the drift parameter θ (see Section F.2 for details). Based on [Lakens, 2014a, Dupont, 1983, Proschan et al., 2006], we implemented the remaining functions in R to enable to automatic calculation of the adjustments.

4.5 Demonstration

In Section 4.3, we outlined three use cases where SPEED would be particular useful to address the challenges that researchers face when planning the sample size. In this section we illustrate the value of using SPEED for the large online study by Hofman et al. [2020] (use case 1) and for a replication of a small-sample study (use case 3). For the latter, since the data for Mathis et al. [2021]'s study is not publicly available, we use instead Smart et al. [2020]'s study, which enables us to explain the entire process including planning the sample size with BUCSS based on a previous small-sample study. For use case 2 (studies with hard-to-access participants), the benefit of stopping early is obvious and we do not go into a detailed example.

4.5.1 Large Online Studies

In the use case of Section 4.3.1, we assume that Hofman et al. [2020] conduct an *a priori* power analysis that suggests a sample size of 1,700. We also assume that the authors learn from the pilot study that roughly 30% of the participants are excluded as they fail the attention-check task. Thus, the authors plan to recruit a maximum of 2,400 participants to obtain 1,700 valid samples. We plan to conduct two interim analysis after 800 and 1350 valid data points. To create the interim analysis plan, we use the exponential spending function with $\nu = 0.3$ for a relatively constant trend. After recruiting 1,085 participants, we have 800 valid responses, and have thus reached the first interim analysis stop. All four p-values are smaller than the nominal α, which allows us to stop the data collection and calculate the adjusted results.

This example shows how SPEED can help researchers plan an online study with a larger sample size for to ensure data quality, yet stop early if the effect is larger than expected. With SPEED, researchers can address the challenge of recruiting additional participants without the risk of spending unnecessary resources.

Interim Analysis Plan & Results : Interim Analysis Plan

	N	nominal α	actual N	unadjusted p-values H1.1	H1.2	H2.1	H2.2
Interim Analysis 1	800	.0196	1,085	2.02E-7	2.28E-4	9.37E-33	8.59E-3
Interim Analysis 2	1,350	.0188	–	–	–	–	–
Final Analysis	1,700	.0116	–	–	–	–	–

Effect Sizes & Adjusted Results SESOI: 0.068

	Cohen's d [95% CI]	adj. p-value†	adj. mean difference	adj. 95% CI
H1.1	0.52 [0.32, 0.71]	2.02E-7	19.91	[10.67, 23.60]
H1.2	0.38 [0.18, 0.59]	2.28E-4	14.92	[5.69, 18.61]
H2.1	1.27 [1.06, 1.47]	9.37E-33	58.44	[75.96, 88.88]
H2.2	0.27 [0.07, 0.48]	8.59E-3	11.44	[2.20, 15.13]

† Adjusted p-value can be compared with the conventional α = .05.

Figure 4.5: A demonstration of SPEED based on Hofman et al. [2020]'s study. The plan (on a grey background) could be preregistered. The interim analysis is based on the exponential spending function with $\nu = 0.3$ for a relatively constant trend. The data collection can stop after Interim Analysis 1 as all p-values are statistically significant.

4.5.2 Replicating Small-sample Studies

We use SPEED to demonstrate the planning and analysis of a visualization experiment that compares algorithms for generating color ramps [Smart et al., 2020] (N = 31). We chose this study because the sample size could have been informed by previous work [Correll et al., 2018], and both papers have made their data publicly available. For demonstration purposes, we assume that the maximum possible sample size is 40 participants, and resample their dataset. For the SED parameters, we follow the recommendations in Section 4.4.2, Box 2, and Box 3. Supplementary Material S1 provides reproducible R code for resampling, planning, and analysis.

In this demonstration, we use only the K-MEANS condition, which is their novel technique, and the LINEAR condition, which is the baseline. Their experiment also uses three visualization types. Hence, we analyze it with a 2×3 repeated-measure ANOVA before running a post-hoc test for the hypothesis. The dependent variable is the er-

ror in an identification task. To plan this experiment, we use a result from [Correll et al., 2018] that compares continuous v.s. discrete color ramps: $t(23) = 4.09$. Based on this input, BUCSS determines that the sample size at 80% assurance level is 25. Since the study assigns experimental conditions in random order, any sample size can be considered balanced. Therefore, given the maximum sample size constraint mentioned above, we plan SED at three equal intervals: 25, 33, and 40. SESOI was also calculated based on N = 40. These pieces of information, which could be preregistered, are shown in the grey-background part of Figure 4.6.

At the first interim analysis (N = 25, Figure 4.6a), the p-value is higher than the nominal α. (In fact, the ANOVA does not find the effect of the algorithm to be statistically significant either.) Since Cohen's d is still higher than the SESOI, the data collection must continue. At the second interim analysis (N = 33, Figure 4.6b), the p-value is lower than the nominal α. Therefore, the data collection can stop and the adjusted results are calculated (Figure 4.6c). For comparison, we also calculated the results as if this experiment had been run with a fixed-sample design (Figure 4.6d). In this example, using SPEED for this experiment would have saved seven (18%) participants while yielding a similar confidence interval estimate. Appendix G presents the results of running the same simulation 1,000 times, showing that SPEED reduces the number of participants by 19% on average for this experiment.

4.6 Checklist for Authors and Reviewers

To encourage rigor and transparency in adopting the methods we presented in this paper, we propose the following checklist for authors. Reviewers can also use this checklist to asses rigor and transparency of the practice and constructively critique the work. If the publication outlet supports multi-stage review, e.g., registered reports—where the methods are reviewed prior to the data collection, the checklist in the **Planning** section could be considered a reviewing criteria.

Planning

☐ 1. *Select the primary hypotheses that are used for planning the sample size.* For clarity, we recommend formulating these hypotheses as

Null hypothesis: There is no difference in error between the K-MEANS and the LINEAR condition			SESOI ±0.07			
	N	nominal α	p-value	Cohen's d	Decision	
Interim Analysis 1	25	.0055	.007346†	0.24	Continue	Ⓐ
Interim Analysis 2	33	.0174	.000096	0.30	Stop	Ⓑ
Final Analysis	40	.0271	–	–	–	

SED decision rules: Stop when either:
- p-value < nominal α → statistically significant
- Cohen's d < SESOI → futility

† Omnibus test also not statistically significant

: Planning information that can be preregistered

	N	α	p-value	Cohen's d [95% CI]	
Results (SED)	**33**	**.05**	**.00720**††	**0.30 [0.14, 0.46]**	Ⓒ
Results (Fixed-sample)	40	.05	.00032	0.26 [0.11, 0.41]	Ⓓ

†† Adjusted p-value can be compared with the conventional $\alpha = .05$. Unadjusted p-value and nominal α are in row B.

Figure 4.6: A demonstration of SPEED based on Smart et al. [2020]'s study. The plan (on a grey background) could be preregistered. The interim results (A, B) enable an early stop. The final result (C)—using seven participants fewer than in the original experiment—has a similar confidence interval as the fixed-sample design (D).

null hypotheses. It is important to state whether each hypothesis is two-tailed (e.g., "no difference between A and B") or one-tailed (e.g., "A is not higher than B") because these choices influence SED and BUCSS calculations and how the SESOI should be compared with the interim effect sizes.

☐ 2. *Justify the choices of power analysis method and effect size estimates.* If the results of previous studies from the literature are used as a basis, cite the specific source experiment(s)—because one paper may contain multiple experiments. Provide relevant statistical values from the literature in an appendix. If additional processing were applied to these results, share the code. When no literature effect sizes are available, explain how the effect sizes were estimated, e.g., whether they were based on pilot studies or researchers' own estimates. Based on these pieces of information, select an appropriate power analysis method (Section 4.4.2), and state which one was used with the respective parameter settings. If ranges of parameters were explored [Wang et al., 2021], provide bounds in the supplementary materials.

☐ *3. Justify the choices of the full sample size (N) and intervals for the interim analyses.* Both N and the intervals should yield balanced data. For example, in a 2 $\times$ 2 Latin-square design, they should be divisible by 4. Explicitly state whether N was decided based on an estimate of an upper-bound effect size or based on researchers' resource constraints. If researchers can afford higher N than suggested by the power analysis, consider trying BUCSS with a higher level of assurance. To determine the first interval of the interim analysis, researchers can use a sample size from similar small-sample studies or a local standard of the application domain. The subsequent intervals can be spaced unevenly if justifiable by other practical constraints.

☐ *4. State the level of assurance for BUCSS and the spending function for SED.* See section 3 for a discussion on the characteristics of these choices. We make general recommendations for HCI in Box 2.

☐ *5. If multiple hypotheses are used to plan SED, choose a method to adjust p-values for multiple comparisons.* Multiple-comparison adjustment methods can be applied either to nominal α or the resulting p-values. We recommend applying them to the p-values to allow the nominal α to be traceable back to the SED parameters. Chen et al. [2017] provide a summary of the methods, which can be helpful for choosing and justifying them.

☐ *6. Preregister.* Preregistration provides a record that a study is preplanned with SED instead of adding SED after data collection has started. Ideally, the answers to all the points above should be preregistered. SED preregistration must include: (1) the primary hypotheses, (2) SESOI, (3) N and analysis intervals, and (4) the nominal α for each interval. See Figure 4.6 for an example.

Data collection

☐ *7. Withhold the results from the interim analyses from the experimenters.* Knowing the exact results from the interim analysis may change how the experimenter behaves in further data collection. To minimize this potential experimenter bias, a separate person could run the interim analyses and let the experimenter know whether to continue or stop the experiment [Lakens, 2014a]. The experimenter can, however, prepare the analysis code such that it only

needs to be executed during the data collection. This could be even more improved if the experimenter is not the same person as the one who designed the experiment.

Statistical analysis

☐ *8. Record and provide the results of the interim analysis and decision made at each interval.* The interim results should be provided as supplementary material to the publication. For readers' convenience, provide a short summary of decision rules in the vicinity as shown in Figure 4.6.

☐ *9. Run other statistical analyses only after data collection has stopped.* SED controls Type I errors during the interim analyses only for the primary hypotheses that are used in the planning. Running other statistical analyses during interim analyses will increase the probability of Type I error.

Reporting

☐ *10. Report the unadjusted and adjusted p-values* The literature on SED provides arguments both for and against adjusting the p-values of the final analysis [Proschan et al., 2006, Pocock, 2005, Dupont, 1983]. We recommend reporting the adjusted p-values in publications because (1) it could be interpreted with respect to the conventional $\alpha = .05$, and (2) unadjusted p-values should have been provided as the result of the last analysis in supplementary material according to our recommendation above. If the authors choose to use the unadjusted p-values in the publication, they must clearly state the nominal α—which is lower than .05—as a reference point.

4.7 Limitations of SPEED

4.7.1 SED and Bayesian Analysis

We presented SED in the frequentist statistics paradigm—and in both the null-hypothesis significant testing and estimation approaches. SED experiments can be subjected to typical Bayesian analysis without any changes, only the planning step requires extra work. For readers interested in applying SED to Bayesian analysis, we suggest [Jennison, 1999, Chapter 18] and the recent update by Schönbrodt et al. [2017]. Future work should extend the R template to Bayesian analysis and planning. It would also be interesting to compare the characteristics of frequentist SED v.s. Bayesian SED in a simulation study based on data collected in the field of HCI.

4.7.2 Using SPEED with Ordinal and Categorical Dependent Variables

Many measures collected by HCI experiments are ordinal, e.g., Likert-style questions, or categorical, e.g., choosing the most preferred item. If they are not used to plan the experiment, they can be analyzed after stopping data collection. To plan SED with categorical dependent variables, we refer readers to [Jennison, 1999, Chapter 12].

Research and software packages for planning SED with ordinal dependent variables are limited. Pocock [1977] presented a small simulation with the Wilcoxon test. However, several works [Pocock, 1977, Mehta et al., 1994, O'Brien and Fleming, 1987] point out that planning using a SED procedure with rank-tests may reduce statistical power, i.e. it is more difficult to get statistical significance. Mehta et al. [1994] present an algorithm using exact permutation tests to address this problem. To the best of our knowledge, no software package yet incorporates this algorithm nor addresses SED planning with ordinal dependent variables. Another direction for future work is to create usable user interfaces for resampling and permutation methods.

4.8 Web application for exploring sample size decisions with SPEED

The SPEED protocol makes several decisions on effect sizes and sample sizes explicit. This change creates an opportunity for software to support users in interactively comparing how each decision influences the candidates for experimental designs. In this section, we present the major decisions in terms of data and task abstractions and describe the state of the existing software tools. We present a web application SPEEDX that is designed to support experienced empirical researchers who are new to the techniques in SPEED protocol. Last, we evaluate the application with the Cognitive Dimensions of Notation framework.

4.8.1 Data Abstraction

The data and task abstractions enable us to compare existing visualization techniques without discussing domain-specific details. We use the What-Why-How framework [Brehmer and Munzner, 2013, Munzner, 2015]. In the following, all items and attributes are quantitative unless otherwise indicated. For the SPEED protocol, two data abstractions are of particular interest:

Sample size tables

An experimental design *candidate* is derived from counterbalancing constraints, Type I error, and either an effect size (in *a priori* power analyses) or a fixed level of statistical power (in sensitivity power analyses). Each candidate is represented as an abstract data table consisting of data items (rows) and attributes (columns). Each item is a sample size. The first attribute (column) depends on the type of power analysis: it is either statistical power levels (for *a priori* power analysis), assurance levels (for BUCSS), or detectable effect sizes (for sensitivity power analysis). The second attribute is categorical values indicating whether a sample size is selected or not. The largest selected sample size is the final sample size. The remaining selections are interim analyses.

The user may come up with multiple candidates for experimental designs. They could have different effect size estimates or different counterbalancing designs. Different effect size estimates yield different values for the first attribute. Different counterbalancing designs impose different constraints on the valid sample sizes—the items. For example, a design with total random assignment can use any sample size, whereas a within-subjects experiment with two conditions requires the sample sizes to be multiples of 2 to be fully counterbalanced. Therefore, two candidates may form a data table with different lengths. For these reasons, each candidate design is represented as an individual abstract data table; multiple tables could be used simultaneously. In theory, the sample sizes can be any positive integer; therefore, the abstract tables have enumerable infinite items. In practice, resources for experimental design are always limited, yielding finite tables.

4.8.2 Task Abstraction

Each task abstraction consists of an `action` and a `target`—which can be thought of as a verb and a noun. Since the procedures in SPEED are likely to be unfamiliar to the readers, we describe each task by giving the abstract `action` first with the concrete targets and then summarizing the `action`-`target` pairs at the end. While planning a SPEED experiment, users have to perform the following two tasks.

T1: Selecting when to conduct interim analyses

Wang et al. [2021] previously present task abstractions for determining a *single* sample size from *a priori* power analysis (T3 and T4). We expand on Wang et al. [2021]'s task abstractions to cover selecting *multiple* sample sizes in sequential experimental design component of SPEED.

A user starts with knowledge about their a maximum affordable sample size ($s_{\max}$). This number may be a precise number or a rough range. Then, the user will `locate` the minimum acceptable sample size ($s_{\min}$)—which is likely based on the convention of their field of study. All lower sample sizes will not be considered.

In an *a priori* power analysis, they will also `locate` s_p: the smallest sample size that has an acceptable statistical power—say, 0.8. If the resources allows $s_{\max}$ being higher than s_p, the user will `browse` the sample sizes within the *range of interest*, $[s_{\min}, s_{\max}]$, where $s_{\min} < s_p < s_{\max}$. One or more sample sizes within this range will be `annotated` as either interim analyses or final sample sizes. The final sample size is likely to be higher than s_p, but the interim analyzes can be anywhere within the range of interest. The sample sizes with the power below s_p allows for a possibility that the population effect size turns out to be higher than anticipated. To make these choices, the users will also consider the concave relationship between power and sample-size: At low sample sizes, the power increases rapidly until passing an inflection point. Then, the increments slow down until eventually plateau. Therefore, the `browse` action incorporate several subtasks: The user will `compare` the difference in statistical power between two or more sample sizes and consider whether that power differences worth the investment of their resources. These `browse`, `compare`, and `annotate` actions generally proceed iteratively from low to high sample sizes. The task abstraction for *a priori* power analysis above also applies to BUCSS because the relationship between sample sizes and assurances exhibits the same characteristics as those between sample sizes and power (Section 4.4.2).

When s_p is unattainable due to resource limitations, the range of interest becomes $[s_{\min}, s_{\max}]$, where $s_{\max} < s_p$. If $s_{\max}$ yields far too insufficient power, the SPEED protocol suggests changing to a sensitivity power analysis. Here, the relationship between the sample size and the effect size is a convex, monotonically decreasing function. Therefore, the `browse`-`compare`-`annotate` iterations are likely to proceed in the opposite direction: from high to low sample sizes. Additionally, the user can also consider hypothetical gains of investing more resources into the experiment. Such consideration could lead to increasing their upper limit $s_{\max}$ and restart the iterations for some or all candidates.

Since the user may consider multiple experimental design candidates, the `compare` actions above may be performed across candidates. Additionally, when design candidates are generated with different effect size estimates, the users may `compare` the overall trend between candidates to assess their differences in the rate of trade-off. These comparisons are likely to be limited to the sample sizes within the range of interest. For *a priori* power analyses, candidates with higher overall slope are favorable because a small increment of the sample size

yields higher power gain. In contrast, for sensitivity power analyses, candidates with lower overall effect size values are favorable.

In summary, the task of selecting when to perform interim analyzes can be abstracted into five abstract tasks: `locate`, `browse`, and `annotate values`, `compare features`, and `compare trends`.

T2: Choosing a Spending Function Configuration

Users have to choose a spending function and its parameter to assign the nominal alpha values to each interim analysis stops. Users select a spending function by name and then *discover* the *trend* of the spending function. Then, users *compare* the *trend* of spending functions with different parameter settings to *identify* which parameter best suits their expected results (Figure 4.3). SPEEDX includes a new mapping algorithm (see Section 4.8.4) that allows users to select a *trend* for the spending function rather than a parameter. A concrete example is that a monotonically decreasing spending function would be preferred when a large effect size is expected, which, in HCI, could be the case if a new interaction technique or artifact presents a major leap over current work.

4.8.3 Current Systems

Three GUI applications currently exist to support sequential experimental design: GroupSeq [Pahl, 2018], gsDesignExplorer [Anderson, 2020], and RPACT [Wassmer and Pahlke, 2022]. gsDesignExplorer and RPACT provide more advanced features beyond the SPEED protocol. We will discuss only the characteristics relevant to the tasks identified in the previous section. Additionally, none of these applications incorporate *a priori* power analysis or sensitivity power analysis.

For **T1**, none of the application supports iterative `browse`-`compare`-`annotate` actions. Instead, all three applications asks the user to specify when to conduct interim analyses as fractions of the final sample size (e.g., 0.5 for half of the total sample size). Users must manually consider counterbalancing constraints to avoid conducting interim analyses when the sample is not fully counterbalanced.

In GroupSeq and RPACT, users select a spending function and specify its parameter. The gsDesignExplorer includes an extra step where users first need to select a spending function family before selecting the function itself. In GroupSeq and gsDesignExplorer, the spending function chart does not include the nominal alpha values, but only shows z-scores. To review the nominal alpha values, the user has to manually convert them with an external tool. Lakens [2014a] created a detailed tutorial that outlines these steps with Microsoft Excel. Exploring different spending function configurations requires multiple steps, and thus remains cumbersome (Task 2).

GroupSeq is a standalone application that is launched from the R console, whereas the gsDesignExplorer and RPACT are R shiny applications for the web. During the exploration process, GroupSeq uses one window as input and for each calculation a new output window. While some input parameters are available in the output window, e.g., the spending function, others are missing, e.g., the input parameter to the spending function. This way it remains challenging to compare different designs as the user needs to keep track of the parameters manually. Due to the nature of being a web application, multiple versions of the gsDesignExplorer and RPACT can be opened side-by-side to explore different design decisions. None of the application facilitate a direct trade-off comparison between several design alternatives (Task 3).

Ultimately, gsDesignExplorer and RPACT were designed for an experienced user group that make use of more complex experimental designs.

4.8.4 Interaction Design

The goal of SPEEDX is to facilitate the exploration and design of group-sequential designs. The user interface consists of of four different panels. Users use the inspector panel (A) to specify parameters such as statistical power or the spending function configuration. They can observe the power analysis result in (B) and set the sample size for the interim analyses. (C) shows the spending function configuration in detail and (D) summarizes all design alternatives for comparison. We describe the user interface for specifying and comparing sequential design alternatives according to diverse criteria, e.g., maximum sample size, analysis stops, and spending function configuration. We

Figure 4.7: SPEEDX interface: (A) Users specify parameters about the experiment design, statistical criteria, and spending function in the inspector panel; (B) Users select when interim analyses take place based on the power analysis result; (C) Users inspect the spending function configuration that assigns the nominal α to the interim analyses; and (D) Users save and select design alternatives and variations. Instead of choosing a power analysis directly, users are guided through a decision tree (left) that determines the appropriate power analysis.

continues to highlight the "How" element of the What-Why-How framework [Brehmer and Munzner, 2013, Munzner, 2015] by using the small-capital typeface to indicate VISUALIZATION IDIOMS.

Inspector Panel: Experiment Design

In (A-1), users can alter all parameters that pertain directly to the design of the experiment itself. Users need to specify two parameters that are based on the counterbalancing design. The multiple which is the number for which the counterbalancing is valid. The replications which is the number of conditions a participant does back-to-back so that the researchers can aggregate the measurements.

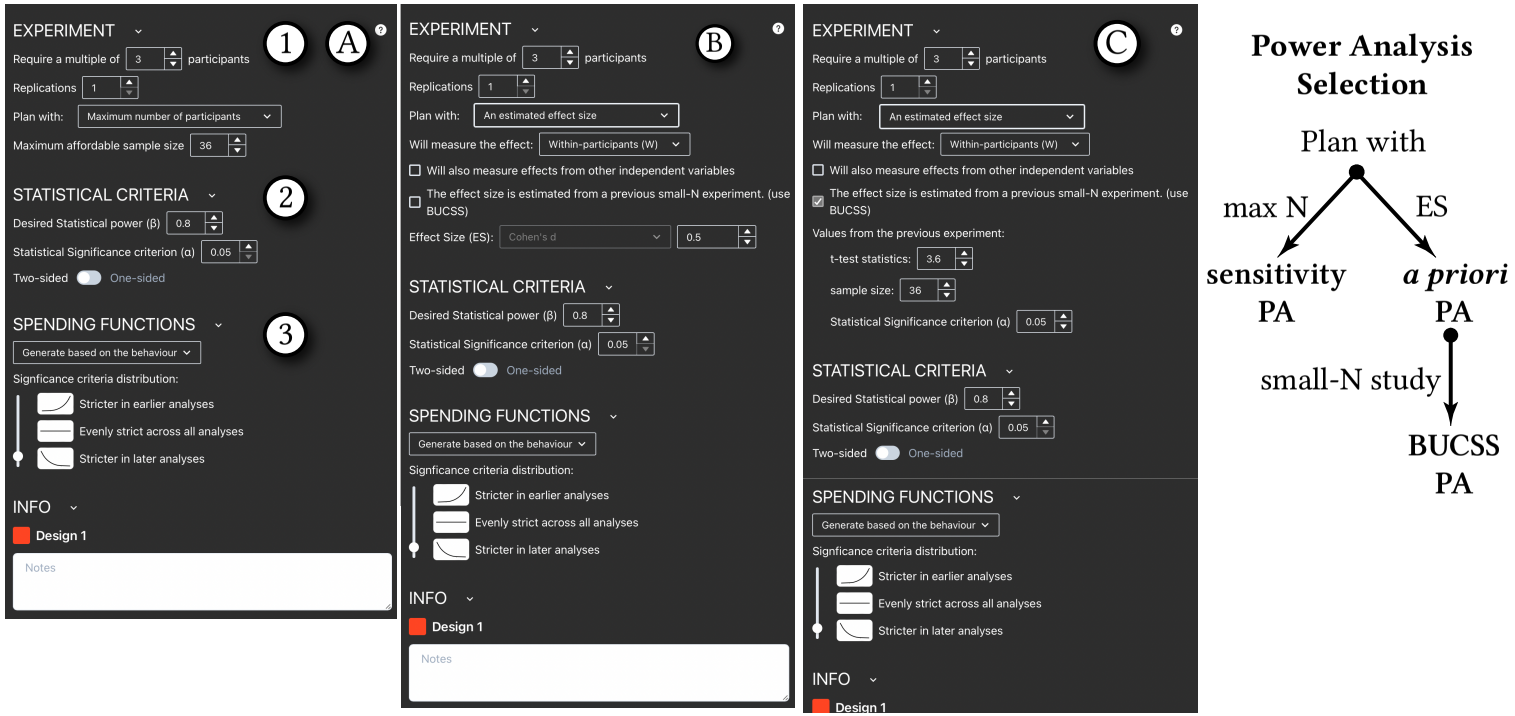


Figure 4.8: Users choose the type of power analysis by selecting "Maximum number of participants" (A) for a sensitivity power analysis or "An estimated effect size" for an *a priori* power analysis (B). To include the BUCSS adjusted, users check the option for "previous small-sample study" (C).

The user also selects a power analysis to specify the maximum sample size and SESOI. First, the user selects whether s/he wants to plan with a maximum sample size directly or an estimated effect size. The former selection shows the sensitivity power analysis where the user species the maximum affordable sample size. When the user selects "plan with estimated effect size", the UI updates to show the a priori power analysis parameters. By selecting the checkbox for previous small-sample experiments, users can switch between the conventional a priori power analysis and BUCSS.

Inspector Panel: Statistical Criteria

Under the statistical criteria, users are able to specify the desired statistical power, the significance criterion (commonly .05), and the directionality of the test.

Inspector Panel: Spending Function Selector

Users can select the spending function configuration either by the function's behaviour or by manually specifying parameters (Figure 4.8-3). The interface presents the behaviour selection first as this requires less experience with spending functions. With a vertical slider, users are able to choose the overall shape of the function based on the exponential spending function. This reduces the time and effort to understand how setting the parameter would influence the nominal alpha values at each interim analysis, thus, facilitating the choice of a spending function configuration (Task 2). When users select the latter option, they are able to choose a different spending function and manually enter the input parameter. This is useful if the user intends to replicate a pre-existing configuration. No matter how the user chooses to change the design, any changes are immediately reflected in the spending function chart on the right.

To enable users to choose the shape of the spending function directly, we first calculate weights that approximate the desired shape before we compute the spending functions with their parameter ranges. The vertical slider ranges from stricter in earlier analysis (1) to stricter in later analysis (-1), where 0 marks evenly strict across all analyses. The

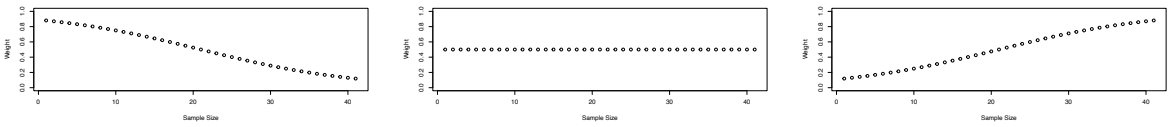


Figure 4.9: The weight profile for the slider that enables users to select the shapes of the spending function directly instead of specifying the parameter manually.

weights are calculated using the following sigmoid function:

$$\frac{1}{1 + \exp(-x \times b)} \tag{4.2}$$

where x is information time at which the interim analysis takes place rescaled to -10 and 10, and b is the strictness level from -1 to 1. Figure 4.9 shows the weight profile at three strictness levels.

After computing the weight profile, we compute the spending function at different parameter and multiply the nominal alphas with the respective weight of the profile. Then, we minimize the differences between all values to select the parameter for the shape that would fit best.

Inspector Panel: Info

Users can give the design a name, adjust its color, and add a comment. This comment can be used to capture design rationales when exploring alternatives.

Power Analysis Chart

Users select when to conduct interim analyses (**T1**) on the Power Analysis Chart. We build upon the visualization in Touchstone 2 [Eiselmayer et al., 2019] as shown in Figure 4.10 (left): It is a line chart with the sample size on the horizontal axis and the statistical power on the vertical axis. Counterbalancing design information is used to FILTER out the sample sizes that are not fully counterbalanced. Without further modifications, this chart already supports the `locate`, `browse`, and `compare` actions in **T1**, but only for single candidate.

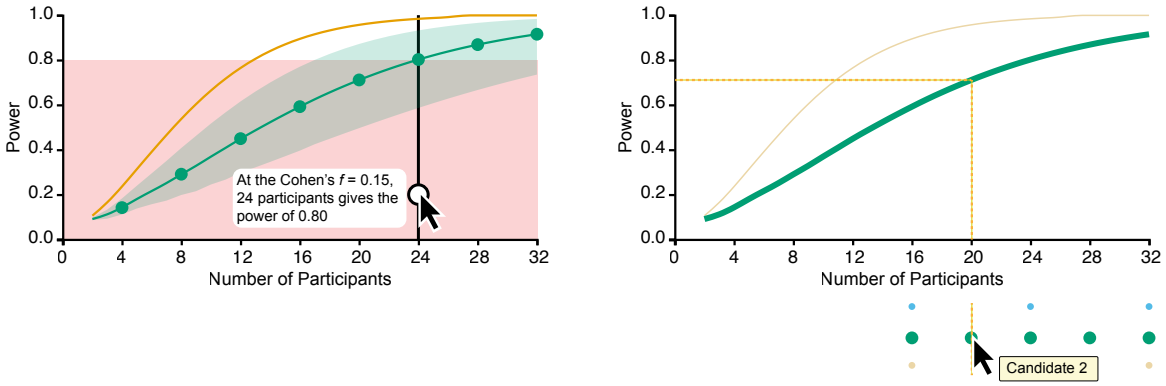


Figure 4.10: Left: The power chart from Touchstone2—adapted from the Figure 1 [Eiselmayer et al., 2019] with permission from the authors to match the current version of their software. Right: SPEEDX power chart.

For multiple candidates, Touchstone2 SUPERIMPOSES multiple curves. An active candidate is highlighted with a confident band calculated from a margin of effect sizes. On the active candidate, fully-counterbalanced sample sizes are encoded in with point marks directly on the curve. The user selects the sample size by moving a knob that is coordinated with a horizontal line that spans the entire chart. Finally, an area with too low power is indicated by a red area mark on the background.

Supporting **T1** in the SPEED protocol is more difficult because the user will select multiple sample sizes for each candidate. Furthermore, several candidates may have identical curves overplotted at the same location while having different sets of selected sample sizes.

The SPEEDX chart is shown in Figure 4.10 (right). The user switches between candidates by selecting one of them in the Overview Table—which will be described later. The active candidate is plotted with a thicker line, and inactive candidates are plotted with reduced color saturation. Since the users can express the uncertainty of the effect sizes by creating multiple candidates, the confidence band is no longer necessary and is removed. We also remove the point marks that indicate fully-counterbalanced sample sizes. Instead, moving the mouse cursor in this chart shows a crosshair that snaps to the nearest fully-counterbalanced sample size of the active line and its corresponding power values. The color of the cross hair is defaulted to the same as the active line. But when the mouse cursor hovers on the sample size that will yield low power, the crosshair changes to yellow to warn

the user of a low power; this sample size may be used as an interim analysis, but it should not be used as the final sample size. This interaction technique replaces the red area mark in Touchstone2. These design choices allow users to `compare values` and `trends` across the curves that are not overplotted.

To facilitate the `comparing` the candidates with totally overplotted curves, below the horizontal axis SPEEDX adds the *sample size indicator*: rows of one-dimensional dot plot—one row per candidate experimental design. Each dot represents a sample size selected for an interim analysis, and the rightmost dot of each row automatically becomes the final sample size. The active candidate row has their dots bigger than the others. The tracking line from the main chart continues into the sample size indicator, and its horizontal position is ALIGNED and SYNCHRONIZED. Hovering over each dot reveals a tooltip showing the name of the design candidate. The sample size indicator substitutes SUPERIMPOSITION with JUXTAPOSITION.

To `annotate` sample sizes, the user clicks anywhere on the power chart to toggle between selected/unselected. The user can also click on the sample size indicator area to toggle any sample sizes, including those from the inactive candidates. Unlike in Touchstone2, SPEEDX prohibits selecting the sample sizes that are not fully counterbalanced. This design decision is necessary to constrain the selections to be meaningful for the subsequent **T2**. Clicking on the sample sizes that are not fully counterbalanced will result in a beep and an explanation message in a tooltip.

The description above applies to the *a priori* power analysis mode. For other modes, the vertical axis changes to the assurance for BUCSS, and the detectable effect sizes for sensitivity power analyses. The visualizations and user interactions remain the same.

Spending Function Chart

The purpose of the Spending Function Chart (Figure 4.11) is to select and compare spending function configurations by comparing the overall shape as well as the nominal alpha values at each analysis (Task 2). The horizontal or x axis shows sample sizes at valid multiples which is the same as in the power analysis chart. The vertical or

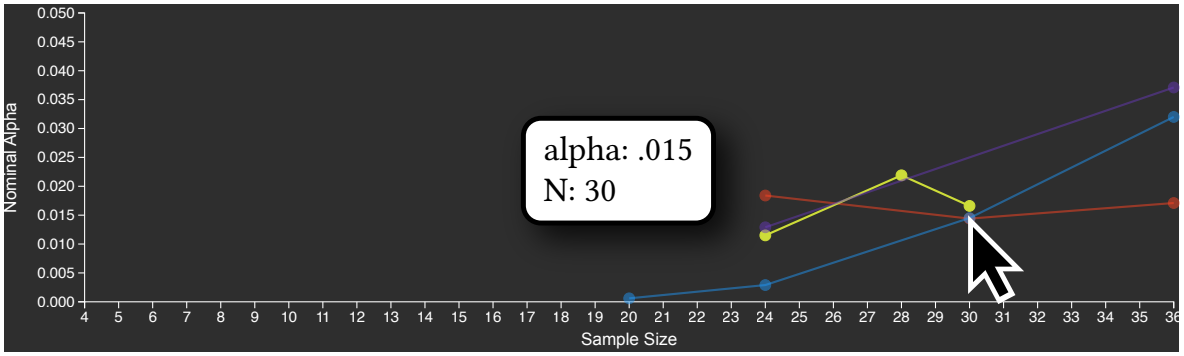


Figure 4.11: The spending function configurations of different design alternatives are superimposed to make the facilitate their comparison.

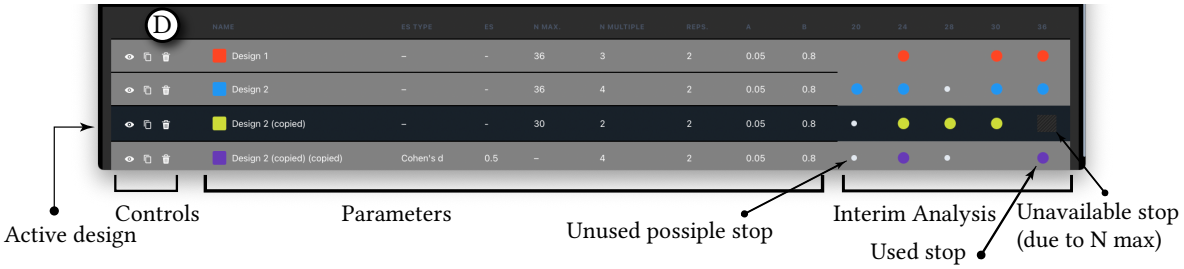


Figure 4.12: The overview table facilitates the comparison of design alternatives. Additionally, users can duplicate, delete, and hide designs.

y axis shows the range of nominal alpha values from 0 to the significance criterion set in the inspector, conventionally .05.

The spending function chart shows the current active spending function with full opacity at the top, while others' opacity is reduced to clearly distinguish the different functions (Task 3). Each function is colored according to the design color. When the mouse is moved over the visualization, a small pop-up window follows the cursor, indicating the current sample size as well as the nominal alpha value at the sample size.

Overview Table

The purpose of the table is to give an overview across the designs and let users identify how the alternatives differ (Figure 4.12, Task 3). Each top-level row represents one design that unfolds to reveal variations

that users can save during the exploration process. The whole row consists of three sections: the controls for the design, the parameters of the design, and the interim analysis stops.

The controls allow the user to hide a design which excludes it from the two visualizations. Additionally, users can duplicate and delete a design. By saving a variation, users create a snapshot of the current state of the design to revisit at a later stage. Users can revert to any variation.

The sub-table storing the information for the interim analyses uses the sample sizes where at least one design has a planned interim analysis. Each cell contains one of the following four codes:

- a colored dot: this indicates that an interim analysis is planned at the given sample size;
- a small grey dot: this indicates that no interim analysis is planned but would be possible at the given sample size;
- a striped pattern: this indicates that an interim analysis is not possible because it exceeds the maximum sample size of that particular design; and
- an empty cell: this indicates that no interim analysis can be planned at the sample size because it halides the multiple constraint.

4.8.5 System Architecture

SPEEDX is implemented as a web application using NextJS Typescript and a relational database with Prisma.io. The statistics are computed using R and served as an API server with R plumber.

4.8.6 Evaluation: Cognitive Dimensions of Notation

In this section, we evaluate the usability of SPEEDX and define design implications for future work by comparing it with the existing tools listed in Section 4.8.3. For this purpose, we use the Cognitive Dimensions of Notation (CD) framework [Blackwell et al., 2001, Green

and Blackwell, 1998], an evaluation framework used in HCI and the information visualization community [Gori et al., 2020, Zong et al., 2021, McNutt and Chugh, 2021, Sarma et al., 2021]. This framework provides a vocabulary for assessing the cognitive impact of design decisions.

A cognitive dimension analysis assumes a specific user group and user activities. For the following analysis, we assume that the users are knowledgable in typical experimental designs but are new to the sequential experimental design and to the SPEED protocol. We believe that this user profile describes many HCI researchers. However, we would like to point out that GroupSeq, RPACT, and gsDesignExplorer were probably not designed with this user group in mind. Therefore, the analysis below is not against these tools. Instead, it highlights design characteristics that broadens the target user group.

As for the user activities, we use those described in the task analysis (Section 4.8.2). In each of the subsections below, we recap each user activity before describing the analysis of relevant cognitive dimensions.

We conducted this analysis with all 14 cognitive dimensions from the CD framework, shown in Table 4.6. In seven cognitive dimensions, SPEEDX is on par with the current software; in the remaining dimensions SPEEDX outperforms them. Below, we discuss salient differences in seven dimensions.

Hidden Dependencies, Premature Commitment, and Progressive Evaluation

The differences along the cognitive dimensions of *Hidden Dependencies*, *Premature Commitment*, and *Progressive Evaluation* are apparent when users alternate among counterbalancing design, **T1** (power analyses), and **T2** (choosing a spending function).

If the experiment design process was straightforward, users would first design counterbalancing, then predetermine the total sample size (N), and then choose the points to conduct interim analyses (e.g., $n_1 = 0.25N, n_2 = 0.5N, \ldots$). In actual practices, however, researchers alternate between counterbalancing design and power analysis in several iterations [Eiselmayer et al., 2019, Section 4]. Each iteration could

No.	Cognitive Dimension	GroupSeq	RPACT	gsDesignExplorer	SPEEDX
1	Viscosity				✓✓
2	Visibility	=	=	=	=
3	Premature Commitment				✓✓
4	Hidden Dependencies				✓✓
5	Role-Expressiveness	=	=	=	=
6	Error-Proneness	=	=	=	=
7	Abstraction	=	=	=	=
8	Secondary Notation	=	=	=	=
9	Closeness of Mapping				✓✓
10	Consistency	=	=	=	=
11	Diffuseness	=	=	=	=
12	Hard Mental Operations				✓✓
13	Provisionality				✓✓
14	Progressive Evaluation		✓✓	✓✓	✓✓

Table 4.6: Comparison of GroupSeq [Pahl, 2018], RPACT [Wassmer and Pahlke, 2022], and gsDesignExplorer [Anderson, 2020] with our SPEEDX using the Cognitive Dimensions of Notation framework [Blackwell et al., 2001, Green and Blackwell, 1998]. Checkmarks indicate an improvement over other applications, and equal signs indicate similar quality.

change the constraints the range of interest and the power levels in **T1** and the nominal-αs in **T2**.

Suppose a pilot study shows that an experiment is too long. One way to address this problem is to remove an experimental condition. However, this removal could change the number of participants required to keep the experiment fully counterbalanced. The removal could also change the effect size in an *a priori* power analysis—which also changes the minimum sample size s_{min}. In both cases, the change could influence the cost-benefit assessment of when to conduct interim analyses. For example, if the original experiment requires at least 40 participants, with the possible interim analyses every 10 participants. A researcher may decide to conduct the first interim analysis with 20 participants. Suppose that a revised design is fully counter-

balanced at every three participants. The smaller cost for each increment (of 3 instead of 10) may induce the same researcher to conduct the first interim analysis with a higher number of participants (e.g., 27)—which is likely to yield a higher statistical power.

This example shows how upstream decisions in the counterbalancing design could expand or limit possible choices in sequential experimental design. Since either counterbalancing, power analysis, or both could influence the choices, the dependencies are branching. Additionally, the chain of dependencies could be long (e.g., counterbalancing choice $\rightarrow$ eligible Ns $\rightarrow$ possible interim choices $\rightarrow$ cost-benefit differences). When these branching and distant dependencies are hidden, the users are likely to hope for the best solution rather than thoroughly explore the possibilities [Green and Blackwell, 1998]. This is a problem in the *Hidden Dependencies* cognitive dimension.

The three current software tools ask the user to determine the total sample size (N) and when to conduct interim analyses as the fractions of N. This input method requires the user to manually apply the constraints from counterbalancing to these choices. When the users change their counterbalancing and wish to see how their changes affect sequential experimental design outcomes, they need to manually take notes or work in separate web browser windows, in which they have to enter all parameters from scratch. These cumbersome interactions aggravate the *Hidden Dependencies* problem. In SPEEDX, users can see multiple counterbalancing designs simultaneously superimposed in the same Power Analysis Chart (Figure 4.10)—allowing the user to adjust the parameters and visually compare how the changes affect the outcomes. Such interaction design increases the visibility of the dependencies.

In addition, *Premature Commitment* problems occur—requiring the users to make decisions before providing necessary information. For **T1**, the three current software tools require choosing N simultaneously with choosing when to conduct interim analyses. This requirement forces the users to categorize these points of analysis before knowing how their choices could affect the spending functions. They also need to specify all analysis points at once before seeing how the nominal alphas are distributed. In contrast, SPEEDX shows all points that are fully counterbalanced together with the power curve (Figure 4.10). From these points, the user selects the sample sizes that they wish to perform for either type of analysis. The point with the highest number of participants becomes the total sample size, and the

other points become interim analyses. This interaction design eliminates the need for the user to specify and categorize each analysis point upfront. Additionally, each modification updates the nominal alpha chart (Figure 4.11)—giving feedback for the user for their next decision iteration. This immediate feedback is an improvement in the *Progressive Evaluation* cognitive dimension—which is essential for inexperienced users and helpful for expert users [Green and Petre, 1996, section 5.9].

To summarize, SPEEDX reveals the *Hidden Dependencies* among the parameters of three processes: counterbalancing, power analysis (**T1**), and selecting the spending function (**T2**). SPEED also mitigates the *Premature Commitments* when selecting analysis stops, and let the users *Progressively Evaluate* outcomes of their choices of the interim analyses.

Closeness of Mapping and Hard Mental Operations

After selecting the points to perform interim analyses, researchers choose a spending function and its parameter. In this step, the three applications differ along two cognitive dimensions: *Closeness of Mapping* and *Hard Mental Operations*.

The spending function determines how the overall alpha (Type I error rate) is distributed among the analyses as nominal alphas. Each nominal alpha is lower than the overall alpha, but they do not have to be the same across all analyses. For example, in experiments where the cost for each participant is high, or when the anticipated effect size is large, researchers may prefer a distribution that gives permissive (higher) nominal alphas in early interim analyses. In contrast, for experiments where the effect size is expected to be small, being more permissive at higher sample sizes is more desirable to allow statistical power to accrue.

How nominal alphas are distributed is determined by the family of the spending function, its parameter, and the number and locations of the interim analyses. As previously shown in Table 4.3, each family has its own specific parameter range that distributes nominal alphas differently. For example, setting the parameter value to zero in the Hwang-Shih-DeCani family yields the same level of nominal al-

phas across all interim analyses, whereas in the exponential family zero yields monotonically decreasing trends (early-permissive).

To explore the choices of spending functions, GroupSeq, RPACT, and gsDesignExplorer require the users to know each family of spending function, and how each parameter distributes nominal alphas. Activating this knowledge while considering other design concerns could overload the users' working memory. This problem exemplifies the *Hard Mental Operations* cognitive dimension. GroupSeq and gsDesignExplorer provide a field for the user to enter any number for the parameter. RPACT is more helpful. After the user selected a function family, RPACT shows a slider for the parameter. The range of the slider is limited to the valid values for the selected family. These methods to exactly specify the function family and its parameter are useful when researchers wish to use the configurations established in prior works. SPEEDX also provides the same method as RPACT. However, to choose a spending function in for a new domain—such as experiments in HCI—researchers should be able to focus on how the function behaves instead of how it is called or which parameter value is needed.

Therefore, SPEEDX maps the continuum from early-permissive, flat, and to late-permissive behavior into a single slider. SPEEDX shows all spending function families and associated parameters that fit the desired behavior. The users can then select the family and parameter that are suitable to their experimental context. SPEEDX improves the cognitive dimension of *Closeness of Mapping* by letting the users choose their desired distribution type and get the immediate feedback directly.

Provisionality and Viscosity

After exploring several design alternatives, researchers compare their trade-offs to choose their final design. In this step, the three applications differ along two cognitive dimensions: *Provisionality* and *Viscosity*.

Users can save several alternatives while creating sequential experimental designs. However, researchers have to choose one design that fits their experiment best based on parameters, properties, and characteristics. For example, the user created four alternatives designs (A,

B, C, D) that use two methods of planning the sample size. Two (A, B) are using the sensitivity power analysis with a maximum sample size of 36 participants, whereas the other two (C, D) use an *a priori* power analysis that returns 30 participants. Additionally, designs A and C have three analyses at 18, 24, and 36, whereas B and D have only two analyses at 15 and 30 participants. Now, the user needs to decide which of the four designs will be the design for the experiment.

To compare the four design alternatives, GroupSeq, RPACT, and gsDesignExplorer require the user to save all the parameter settings manually for each design as neither of the applications allows users to store temporary alternatives. This means that users need to compare the parameter settings manually. Specifically, in this example, users have to compare the maximum sample sizes, the sample sizes for each analysis, and their associated nominal alpha values. Keeping track and comparing many numbers individually is cumbersome and error-prone. This problem exemplifies the lack of *Provisionality* in the three applications, which is the lack of saving intermediate designs for comparison. SPEEDX allows users to store different design alternatives and snapshots that a user might encounter that could be useful for the future. Users can switch between the different designs at a moment's notice and can compare the parameter settings either superimposed in the visualizations or juxtaposed in the table.

The complete parameter settings must be updated if users want to revisit any of the previous alternatives. To revert the parameter settings, say, change the power analysis and add the interim analysis stops, users have to do many interface actions, thus, leading to a high *Viscosity*. SPEEDX improves the *Viscosity* by including the intermediate designs into the application.

In summary, SED improves upon GroupSeq, RPACT, and gsDesignExplorer in the dimensions mentioned above by allowing users to explore and compare different sequential experimental designs iteratively. GroupSeq, RPACT, and gsDesignExplorer have a high *Viscosity* and low *Provisionality*, making SPEEDX a valuable and suitable tool for designing sequential experimental designs.

4.9 Discussion

The goal of this work is to encourage researchers to use different types of power analysis to plan their experiments and to give them more freedom and expressiveness choosing appropriate sample sizes. This section discusses the broader implications of adopting SPEED for HCI experiments.

4.9.1 Encouraging explicit and nuanced conversations about sample size

The SPEED protocol offers researchers four ways to be explicit about their sample-size decisions. First, researchers' choice of the power analysis indicates their belief about the provenance of the effect size. For example, consider three situations that HCI researchers usually face when planning experiments: (1) when reliable effect sizes exist in the literature, (2) when they exist but are from small-sample studies, and (3) when no effect sizes are available. The SPEED protocol covers all three. In the first case, researchers can directly conduct an *a priori* power analysis. In the second case, researchers use an *a priori* power analysis with BUCSS. In the third case, researchers use their domain knowledge to come up with a minimum effect size that would be practically significant [Kirk, 1996, Dragicevic, 2016] before proceeding with an *a priori* power analysis.

Second, researchers can discuss how the resource constraints influence the intent of their study. Continuing from the example above, suppose the calculated sample size is far beyond the available resources. Without the SPEED protocol, researchers might abandon the power analysis and revert to using a sample-size heuristic [Eiselmayer et al., 2019, p. 4]. With the SPEED protocol, researchers can describe how the sample size suggested by *a priori* power analysis exceeds their resources and then conduct a sensitivity power analysis that determines the smallest effect size of interest (SESOI) that could be detected with the limited sample size. If the difference between SESOI and the expected effect size is reasonably small, researchers could decide to proceed with their experiment. They can even preregister the plan as a confirmatory analysis. Otherwise, if the difference is large, researchers could also use this result as a justification for either (1) substituting their experiments with alternative validation methods [Greenberg and

Buxton, 2008] or (2) intentionally running a preregistered exploratory experiment [Cockburn et al., 2018]. Should they pursue the latter, they can use interval estimates to discuss uncertainty and nuances in their results [Dragicevic, 2016, section 13.4.5].

Third, researchers can use their resources and the participant pool more efficiently. For example, they can weigh the costs of using an unnecessarily large sample size with the benefit of added statistical power. The cost of excess sample size is high in several situations—e.g., when the participant pool is limited or when participants' time is precious or expensive (Section 4.3.2). In such situations, judicious use of participants is prudent and ethical. Aside from these situations, saving research resources is clearly advantageous. For example, researchers can direct the resources spared to conducting internal replication studies that could further strengthen their scientific claims. Alternatively, they can use these resources to conduct experiments that investigate further nuances and interaction effects. On the benefit side, larger sample sizes give higher statistical power. However, the relationship between sample size and statistical power is a concave function: Initially, power increases quickly but eventually plateaus—resulting in a diminishing return.

An *a priori* power analysis lets researchers choose the sample size only once. If their effect size estimate is close to reality, they may choose a close-to-optimal sample size. Otherwise, they may select a sample size that is too small—rendering their entire experiment futile—or too large—unnecessarily wasting the resources and the participant pool. The sequential experimental design part of the SPEED protocol allows researchers to split this decision into multiple interim points where they can consider the effect size and uncertainty from the data collected at each point thus far. This supports more informed decision-making: If their estimate of the effect size is accurate and the chosen sample size is near the optimal, they can continue the experiment. However, if their estimate is far above the real effect size, the SPEED protocol lets them decide to stop the experiment early and redirect the remaining resources into other more promising experiments. By contrast, if the real effect size is much larger than their estimate, the researchers can stop the experiment early and save the resources. These additional decision points are not *p*-hacking or HARKing—Hypothesizing After the Results are Known—because each interim analysis is tested with a lower nominal alpha than the overall Type I Error. The difference between SPEED and HARKing is further discussed in the next subsection.

Fourth, researchers can incorporate more domain knowledge and judgment into the sample-size decision. This additional information is captured in (1) the choices of when to run interim analyses and (2) the shape of the spending function. For example, researchers may use the local sample size standard [Caine, 2016] to run the first interim analysis and use the upper limit based on resource constraints as the full sample size. As described in Section 4.8.6, if researchers are confident that the effect size is likely to be large, they could allocate higher nominal alphas in early interim analyses. These choices are made explicit at the planning stage, and SPEED encourages researchers to describe and justify them to increase transparency in research decisions.

As the field of HCI matures, empirical studies are likely to hone in on nuances of interaction techniques and other phenomena. The focus on these nuances are likely to have smaller effect sizes than earlier works. The field of HCI is also increasingly working with more extreme participant pools—that warrant judicious involvement. Careful experimental designs and sample size decisions enabled by the SPEED protocol will enable individual researchers and the HCI field as a whole to be more efficient.

4.9.2 Sequential experiment design is not HARKing

Although the SED component of SPEED (Section 4.4) may appear similar to HARKing [Cockburn et al., 2018], it actually encourages researchers to explicitly avoid HARKing. John et al. [2012] surveyed 2,155 researchers in Psychology with an incentive for truth telling about ten HARKing methods. Two of these are relevant for SED:

- **Collect extra data:** Deciding whether to collect more data after looking to see whether the results were significant (57% of the participants in John et al. [2012]'s survey indicated that they had personally done this); and
- **Stop early:** Stopping collecting data earlier than planned because one found the results that one had been looking for (19%).

The process we propose in this paper differs from these two practices because (1) SED requires the full sample size to be specified at the planning stage, and (2) the early stops are decided based on nominal

alphas that are more stringent than the overall Type I error (lower than .05).

Additionally, according to the *endowment effect* [Kahneman et al., 1990], we speculate that researchers place a higher value on the data they have collected than on an identical dataset that they have yet to collect. Without the process we proposed, researchers make decisions whether to commit to a malpractice of collecting extra data when they already possess all data they initially planned (N). The temptation to collect extra data at this point could be aggravated by the endowment effect. On the other hand, researchers who use the process we proposed are confronted with decisions at two earlier points: a sensitivity analysis may have suggested that the limited sample size will be inadequate to detect the effect (0 data points collected) or an interim analysis may indicate that the effect is much smaller than anticipated (n data points collected, where $n \ll N$). The decisions to terminate the study early is likely to be less confounded with the endowment effect because the remaining data has not been collected yet. The process we propose indirectly mitigate the file drawer problem by preventing the underpowered studies from being run (or fully run) in the first place.

However, SPEED is not a panacea. SPEED adds a scenario that a decision to stop early may be alluring: When a SED study yields an interim p-value between nominal alpha and .05, researchers may be tempted to stop and treat the results as if it come from a fixed experimental design (FED). A systematic solution for this SED-to-FED malpractice is a widespread expectation of preregistration from reviewers and publication venues for controlled experiments [Cockburn et al., 2018].

Nevertheless, SPEED provides a procedure to mitigate the consequence of SED-to-FED malpractice. Converting a SED to a FED study at an interim analysis overestimates effect size with underestimates its uncertainty [Lakens, 2014a] (e.g., a larger mean difference with a longer confidence interval). For subsequent groups of researchers, they can use BUCSS to adjust the effect size to avoid planning an underpowered follow-up studies.

4.10 Conclusion

This paper presents SPEED, a protocol for helping HCI researchers conduct experiments with an appropriate sample size, and thus avoid statistically underpowered or overpowered experiments. Although *a priori* power analysis is a well-established method, most HCI researchers lack readily available effect sizes. Researchers may also abandon *a priori* power analysis if the resulting sample size is unattainable with their resources or limited participant pool. We address these challenges with three main contributions.

First, we introduce **the SPEED protocol** that enables principled and nuanced decision on the sample-size decisions. The SPEED protocol has three main components: **sensitivity power analysis**, **bias- and uncertainty-corrected sample size (BUCSS)**, and **sequential experimental design (SED)**. Sensitivity power analysis lets researchers estimate the effect size that can be detected given specific resource and participant constraints. BUCSS also lets researchers conduct an *a priori* power analysis that accounts for the small-sample studies common in HCI experiments. SED lets researchers conduct controlled experiments with interim analyses, thus letting them stop collecting data early if the sample effect is much larger or weaker than expected. We explain the benefits of the SPEED protocol with two examples drawn from HCI studies.

Second, we provide **R templates** and a **checklist** for authors and reviewers to plan, conduct, report, and review experiments designed according to the SPEED protocol. Throughout the paper, we offer specific recommendations to help authors and provide a checklist containing the aspects of a thorough procedure.

Third, we developed **a web application, SPEEDX**, to help researchers who maybe new to sequential experimental design and SPEED. We also the analyzed cognitive dimensions relevant to using SPEEDX, and compare them to three existing software tools for planning sequential experimental design. This analysis suggests that SPEEDX's interaction design can lower the barrier for accessing the SPEED.

As a cornerstone of the scientific method, controlled experiments have contributed to establishing HCI as a scientific discipline. It is now time to adopt state-of-the-art methods, such as SED and BUCSS, and processes, such as preregistration, to enable the next generation of ad-

vances in the field. We look forward to the adoption of these or similar methods in the HCI community and to the development of tools to support them.

Chapter 5

Conclusion

5.1 Contributions

Designing good experiments is challenging but crucial to the credibility of the findings and subsequent research. Researchers encounter this difficulty due to the limited availability of tools for non-experts in statistics. Therefore, this thesis aims to assist researchers in the experimental design process by (1) providing tools that facilitate experiment design and sample size planning, (2) gathering information on the challenges and insights experienced by users of these tools, and (3) presenting a protocol for making more flexible sample size decisions. Each of the three projects concentrates on a vital aspect of the experimental design process.

The first project, called *Touchstone2* (Chapter 2), focuses on highlighting the decisions related to independent variables, blocking, and counterbalancing. It allows researchers to examine trial tables of participants. However, the support for determining the number of participants is limited to simple power analysis and counterbalancing. The second project, called *Argus* (Chapter 3), builds upon *Touchstone2*. The work in *Argus* delves deeper into *a priori* power analysis and provides a tool that assists researchers in selecting an appropriate sample size while considering all the constraints identified in *Touchstone2*. Both *Touchstone2* and *Argus* operate under the assumption of a fixed sample size for experiments. Continuing the work, SPEED and SPEEDX enable researchers to employ more flexible sample sizes. These tools allow researchers to plan the counterbalancing design using *Touchstone2*, esti-

mate a maximum sample size using *Argus*, and determine the interim analyses using SPEEDX. Together, these tools empower researchers to design better experiments by providing them with a comprehensive experimental design toolkit. The following is a list of each project, their contributions, and how they address the research questions.

The article in Chapter 2, which discusses *Touchstone2*, presents four contributions:

1. an empirical interview study that identifies the challenges that researchers face during the experimental design process and how they overcome these challenges (empirical contribution);
2. a web application that enables researchers to examine and compare design alternatives, thereby facilitating the evaluation of trade-offs (artifact contribution);
3. a domain-specific language specifically designed for researchers to describe and share their experimental designs (artifact contribution); and
4. two evaluation studies that demonstrate the efficacy of *Touchstone2* in supporting the design process (empirical contribution).

These four contributions address **RESEARCH QUESTION 1:** How can researchers be supported when designing controlled experiments? The article offers a comprehensive analysis of the challenges encountered by researchers in the field of HCI and other related fields during the process of designing controlled experiments. This article has empirical, artifact, and theoretical contributions.

Touchstone2 incorporates an interactive power analysis chart and a form that enables users to calculate Cohen's f effect size using available data. We discovered that the visual depiction of the power curve was beneficial for users, although understanding the effect size remained a significant obstacle, and the input form was cumbersome. Armed with this empirical understanding, we embarked on enhancing the power analysis procedure through the development of *Argus*.

The article in Chapter 3 about *Argus* includes four contributions:

1. a task analysis that focuses on conducting *a priori* power analyses and examines the challenges that researchers face in this process (artifact contribution);

2. the development of a web application that enables researchers to explore the various factors that contribute to sample size, statistical significance, and statistical power (artifact contribution);

3. a use case that illustrates how researchers can use *Argus* to plan the sample size of their experiments based on previous work (theoretical contribution); and

4. a validation study that demonstrates the insights researchers can gain during the power analysis process (empirical contribution).

These four contributions address **RESEARCH QUESTION 2:** How can researchers be supported when conducting *a priori* power analyses to inform the sample size? The article identifies specific challenges associated with using *a priori* power analysis and provides an application that allows researchers to explore *a priori* power analyses. This article has artifact, theoretical, and empirical contributions.

Argus enables users to compare two scenarios by overlaying different charts and temporarily adjusting static input parameters using the history view. Through this process, we discovered that visual exploration, coupled with closed-loop feedback, enhances the exploration of causal relationships between different sets of parameters. The interaction design of SPEEDX closely adheres to the principles and lessons we learned from our work on *Argus*.

The article in Chapter 4 about SPEED and SPEEDX includes three contributions:

1. a protocol that combines various methodologies, enabling researchers to implement early stopping of data collection in experiments (methodological contribution);

2. R templates, guidelines, and a checklist that assist researchers in conducting and assessing the rigor of experiments (artifact contribution); and

3. an application that enables researchers to plan experiments with flexible sample sizes (artifact contribution).

These three contributions address **RESEARCH QUESTION 3:** How can researchers in HCI utilize a more flexible approach to plan sample sizes for controlled experiments? The article outlines a process that incorporates various power analyses and sequential experimental design. It provides examples, guidelines, R templates, and an applica-

tion to aid researchers in planning and conducting such experiments. This article has theoretical and artifact contributions.

5.2 Discussion

This section explores the integration of each project within the broader context of statistical trends and the potential future implications of this research.

5.2.1 Multiverse Analyses

Multiverse analysis has emerged as a recent trend in research, aiming to demonstrate the robustness of findings, increase transparency regarding researcher degrees of freedom, and highlight potential alternative conclusions [Steegen et al., 2016]. In their study, Steegen et al. [2016] presented different multiverse analyses, each showcasing the variation in analysis results based on different data processing approaches. Additionally, Dragicevic [2016] highlighted how sampling errors can lead to diverse outcomes in seemingly similar multiverse analyses. To facilitate multiverse analyses, Sarma et al. [2021] developed a tool called `multiverse` that utilizes a special markup language, eliminating the need for custom code for each universe. This tool streamlines the generation of extensive multiverse analyses but also introduces new challenges, such as debugging code errors and refining the multiverse to focus on meaningful analyses. Gu et al. [2023] presented the MULTIVERSE DEBUGGER, which simplifies the identification and resolution of issues within individual multiverses, aiding users in running specific analyses effectively. Reporting multiverse analysis results presents its own challenges, prompting Dragicevic et al. [2019] to propose the concept of "explorable multiverse analysis reports." These reports allow users to modify and manipulate decisions throughout the analysis process, providing an interactive environment to explore how these alterations influence the reported results.

Recruiting an appropriate sample size is crucial to avoid sampling errors and ensure reliable research findings. The decisions made by researchers when determining the sample size can also be considered within the framework of a multiverse. In the context of sample size

planning, *Touchstone2* incorporated an uncertainty band around the power chart, visually representing the uncertainty associated with the effect size. This feature allowed users to anticipate potential deviations in the effect size, which could influence the choice of a smaller or larger sample size. However, it's worth noting that the uncertainty band in *Touchstone2* does not align perfectly with the concept of multiverse analysis since it doesn't involve researchers making multiple decisions. On the other hand, *Argus* provides users with the ability to explore the uncertainty of the effect size by incorporating hidden confounds and visualizing simulated outcomes in a multiverse format. This capability allows users to analyze and comprehend the relationship between input parameters and the final sample size, increasing confidence in their decision-making process. While both *Argus* and *Touchstone2* primarily focus on the planning phase of the experiment, the concept of multiverse analysis introduced by Steegen et al. [2016] could be applied to the collected data.

With SPEED, researchers are required to consider and prepare certain aspects of the data analysis during the planning stage. SPEED enables users to make informed decisions about stopping data collection early if the observed effect size significantly deviates from the estimated effect size. By implementing early stopping, the uncertainty in the planned sample size is reduced, as researchers can confidently recruit a larger number of participants, knowing that only the necessary amount is required to achieve meaningful results. During the planning phase, researchers determine a spending function configuration that establishes the stopping criteria for achieving statistical significance at each interim analysis. However, even when the appropriate stopping conditions are met, researchers retain the flexibility to decide whether to halt data collection at an interim analysis. This flexibility in decision-making regarding data collection and stopping criteria presents interesting parameters for the creation of a multiverse analysis, particularly if researchers choose to collect data up to the full sample size.

All three projects primarily center around prevailing statistical practices in which researchers typically conduct and report a single analysis. However, SPEED, with its sequential experimental design approach, introduces additional parameters and decisions into the analysis process at an individual level. As for future work, it would be worthwhile to explore how SPEED and SPEEDX, could be incorporated into a multiverse analysis framework. Investigating the integration of

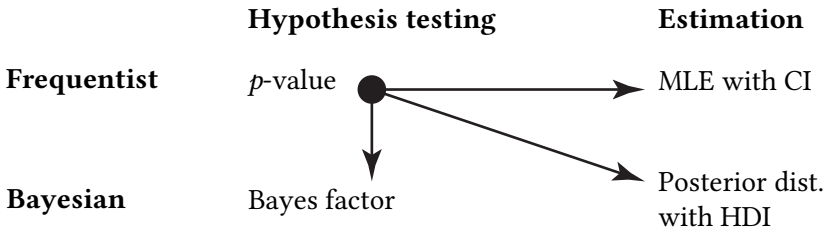


Figure 5.1: This work focuses on hypothesis testing under the Frequentist paradigm. Future work can extend each project to the other types and methods for statistical inference. This figure is based on [Kruschke and Liddell, 2018].

these tools within a multiverse analysis would provide insights into the robustness and sensitivity of research findings.

Touchstone2 and SPEED not only encourage researchers to engage in discussions about experimental design and sample size decisions within the research community but also directly in their own research work. *Touchstone2* offers a powerful declarative language that enables researchers to represent experimental designs and regenerate trial tables using the Touchstone Engine. This feature enhances reproducibility by providing a means to recreate and verify experimental setups. In the case of SPEED, it actively promotes engagement in the sample size discussion by allowing researchers to adjust the sample size based on observed results. This feature necessitates transparency regarding the research plan and analysis, as researchers need to justify and communicate their decisions. Both projects contribute tools and resources that serve as a foundation for future endeavors aimed at improving transparency and reproducibility in research practices.

5.2.2 Types and Methods of Statistical Inference

Researchers can use four different approaches to assess their treatment's effectiveness. Frequentist and Bayesian statistics[1] are two types of statistical inferences, while hypothesis testing and estimation are two statistical inference methods. The matrix presented in Figure 5.1 illustrates the relationships between these different statistical prac-

tices. In the context of this thesis, the focus lies specifically on Frequentist statistics utilizing hypothesis testing.

Frequentist statistics (top row) are based on random sampling distributions that represent different groups within the population. Hypothesis testing (top left) allows researchers to make a binary decision regarding the effectiveness of a treatment. This decision is based on whether the p-value is below a predetermined threshold (commonly .05). However, it is important to note that the p-value only provides information about the presence or absence of treatment effectiveness, without indicating the extent of its effectiveness. To gain a deeper understanding of treatment effectiveness within the Frequentist framework, researchers can utilize estimation-based statistics (top right quadrant) by employing a maximum-likelihood estimator (MLE) along with a confidence interval (CI). The MLE yields a point estimate that represents the most probable center of the data, while the CI indicates the level of uncertainty associated with this estimate. However, the CI does not provide any information about the distribution itself, treating all values within the interval as equally probable as the MLE, according to [Campbell, 2021].

Unlike Frequentist statistics, Bayesian approaches (bottom row) place emphasis on the probability that a hypothesis is true. In Bayesian analysis, the Bayes factor is employed to quantify the strength of evidence in favor of either the null or alternative hypothesis. Unlike p-values, which can be challenging to interpret intuitively, the Bayes factor (bottom left) offers a more straightforward interpretation. It represents the ratio of the likelihoods between competing hypotheses. For example, suppose a researcher's prior is considered an unbiased estimate of their long-term accuracy in selecting hypotheses. In this case, if the Bayes factor is three, it would be interpreted as follows: if the alternative hypothesis is true, the researcher would be correct three times as often compared to if the null hypothesis were assumed to be true. Since the Bayes factor is a ratio between hypotheses, we can not interpret it as a probability. The posterior distribution with the highest density interval (bottom right) represents a probability distribution over parameter values.

[1]Please note that Bayesian statistics refers to Bayesian analysis and not Bayesian Experimental Design (BED) as mentioned in Section 4.2.1. BED is a set of decision-making procedures allowing sample size decisions during data collection. Bayesian analysis utilizes prior information that was available before the experiment to inform the statistical analysis conducted after data collection.

Within each of the statistical frameworks, researchers have two methods at their disposal to draw inferences: hypothesis testing and estimation. With hypothesis testing (left column), researchers strive to reject a null hypothesis that assumes no treatment effect. When the null hypothesis is rejected, researchers can draw conclusions based on significant findings, typically indicated by a significance level such as $p < .05$, suggesting that the treatment is effective. However, the magnitude of the statistical measures, such as *p*-values or Bayes factors, does not provide information about the extent of the treatment's effectiveness. In other words, a smaller *p*-value does not imply a more effective treatment. On the other hand, estimation-based statistics allow researchers to conclude their analysis with a range of plausible values representing the potential effectiveness of the treatment.

The results of Bayesian approaches (bottom row) are often considered more intuitive and easier to understand and interpret, although the analysis process itself may be more challenging. In contrast, Frequentist statistics (top row) have traditionally been the predominant choice for statistical analysis and have remained popular. However, there is a clear trend indicating the growing popularity and adoption of Bayesian statistics in the field of Medicine [Hackenberger, 2019]. This trend suggests that the field of HCI may also follow suit and embrace Bayesian statistics in the future.

5.2.3 Possible Directions for Future Work

This work primarily focuses on Frequentist statistics within the hypothesis testing paradigm (top left) which remains the dominant method for statistical analysis in the field of HCI. In the following discussion, I will explore how each of the three papers fits within the landscape of statistical methods. *Touchstone2* (Chapter 2) uses *a priori* power analysis to suggest an appropriate sample size to the user based on the anticipated effect size and the constraints related to counterbalancing. Although it primarily operates within the framework of hypothesis testing, *Touchstone2* also includes effect sizes for reporting, allowing researchers to incorporate Frequentist estimation (top right). In the context of Bayesian statistics (bottom row), researchers commonly employ "sample size determination" (SSD) to determine the number of participants required for their study [Wang and Gelfand, 2002]. However, *Touchstone2* does not currently support SSD as it requires users to specify various simulation parameters. Neverthe-

less, the aspect of designing counterbalancing remains applicable to Bayesian statistics. Future work could focus on incorporating SSD into the experimental design process of *Touchstone2*. Moreover, there is potential for extending *Touchstone2* by incorporating Bayesian Experimental Design (BED), where the levels of independent variables are informed by previous experiments [Chaloner and Verdinelli, 1995]. In this scenario, users could load or simulate a previous experiment within the workspace to inform the design of a new experiment. While the overall trade-off comparison design may be similar, users would need to select BED parameters to compute a new experimental design.

In (Chapter 3), *Argus* is introduced as a tool that enables users to explore the intricate relationship between parameters and confounding factors within *a priori* power analysis. While the exploration of confounding factors remains a compelling aspect, the specific Bayesian approach of SSD would not be directly applicable in this context. However, it is worth noting that SSD in Bayesian analysis involves an iterative exploration process where researchers experiment with different sample size configurations and prior distributions to simulate multiple posterior distributions [Wang and Gelfand, 2002]. These posterior distributions are then evaluated using performance criteria to inform the specification of a new sample size configuration. This would allow researchers to include the SSD exploration of confounds mainly present in human participants, such as learning or fatigue effects. Future work might examine how a tool similar to *Argus* could facilitate researchers' Bayesian reasoning regarding the sample size.

In SPEED (Chapter 4), researchers compare the p-values of the main hypotheses obtained during an interim analysis with the precomputed nominal alpha value. This approach is compatible with estimation-based statistics under the Frequentist paradigm (top right). Hack et al. [2022] created the R package `AGSDest`, which enables users to calculate the stopping criteria and adjustment for estimation-based statistics, such as effect sizes and confidence intervals. On the other hand, for Bayesian statistics (bottom row), Moerbeek [2021] created a protocol similar to SPEED that uses Bayesian updating. In this protocol, researchers employ the Bayes factor as the stopping criterion at each interim analysis. However, if the Bayes factor fails to exceed the support threshold for either the null hypothesis (H_0) or alternative hypothesis (H_1) before reaching the maximum sample size, the results remain inconclusive. Only when the Bayes factor exceeds either support threshold can researchers report the results, either as hypothesis

testing (bottom left) or estimation-based (bottom right). Future work could explore the extension of SPEEDX to incorporate the use of Bayes factors as a decision criterion.

All three projects have a primary focus on hypothesis testing within the Frequentist paradigm, which is the commonly used statistical analysis practice in the field of HCI. These projects serve as important foundations for supporting researchers in the selection of appropriate sample sizes and also pave the way for the integration of estimation-based and Bayesian statistics. Opportunities for future research include assisting users in determining sample sizes during experiment design, gaining a deeper understanding of the complex relationships involved in planning sample sizes using alternative methods, and expanding flexible sample size planning to accommodate other statistical procedures as they gain popularity.

Within the field of HCI, several methodologies share similarities with or incorporate elements from controlled experiments. For example, in the study conducted by Koch et al. [2020] a structured observation was employed, which can be considered a quasi-experimental approach. Similar to controlled experiments, conditions were manipulated; however, in this case, both quantitative and qualitative data were collected. This approach enhances the ecological validity of the findings, but does not enable causal inferences. The task employed in these methodologies is characterized by a higher degree of realism and less control compared to traditional experiments, resulting in reduced generalizability of the findings. While these methodologies share some similarities with controlled experiments, it is important to note that the focus of this thesis is specifically on unaltered controlled experiments. The tools developed to aid in experiment design have the potential to be applicable to other closely related research methods within the field.

5.3 Closing Remarks

This thesis presents research on supporting researchers during the experimental design process and introduces several tools and artifacts. The research presented in this thesis primarily concentrates on supporting researchers in the field of HCI, but it also has broader applications. The thesis encompasses three projects, each addressing different aspects of the experimental design process:

- *Touchstone2* (Chapter 2) focuses on assisting researchers with the design of counterbalancing, allowing them to determine how experimental conditions are allocated to participants.
- *Argus* (Chapter 3) aims to facilitate sample size planning through the utilization of *a priori* power analysis.
- SPEED and SPEEDX (Chapter 4) provide researchers with the ability to make more flexible decisions regarding sample size.

In conclusion, this thesis makes valuable contributions in the form of artifacts, empirical findings, and methodological advancements to the toolbox of quantitative methods used by researchers. By focusing on controlled experiments, which continue to be a crucial method for collecting empirical evidence, this work holds significant relevance for future research in the field.

Appendix A

Differences in standardized effect sizes formulation

For a **between-subjects design**, we first calculate mean (M_1, M_2) and standard deviation (s_1, s_2) for each group. The simple effect size is the difference between the means, and the standardizer is an average of the standard deviations weighed by the sample size of each group (N_x).

$$d = \frac{M_2 - M_1}{s_\text{p}}, \qquad s_\text{p} = \sqrt{\frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}}$$

Suppose, however, that we **block by handedness** i.e. separating participants into left- and right-handed before randomly assigning each group to the two conditions. The standardizer requires scaling s_p with a factor that excludes the between-block variance s_b:

$$d = \frac{M_2 - M_1}{s_\text{p}\sqrt{1 - s_\text{b}^2/s_\text{p}^2}}, \qquad s_\text{b}^2 = N_\text{LH}(M_\text{LH} - M_\text{all}) + N_\text{RH}(M_\text{RH} - M_\text{all})$$

For a **within-subjects design**, the change happens at the simple effect size. The differences between the two conditions are calculated individually for each participant before being averaged to be the simple effect size (M_diff). The standardizer (s_av) is the average of the standard deviation of the two conditions.

$$d = \frac{M_\text{diff}}{s_\text{av}}, \qquad s_\text{av} = \sqrt{\frac{s_1^2 + s_2^2}{2}}$$

Appendix B

Propagation Algorithm

procedure PROPAGATECHANGE(nodes n, difference d)
    $C \leftarrow$ CHILDRENOF(n)
    $C_u \leftarrow$ UNLOCKEDNODES(C)
    for $c \in C_u$ **do** ▷ Top-down propagation
        $c \leftarrow d \times \|C\| \,/\, \|C_u\|$
    end for
    UPDATE(n)
end procedure
procedure UPDATE(node n)
    $v_{\text{past}} \leftarrow n.value$
    $C \leftarrow$ CHILDRENOF(n)
    $n.value \leftarrow (\sum_{c \in C} c.value) / \|C\|$
    $p \leftarrow$ PARENTOF(n)
    if ISUNLOCKED(p) **then**
        UPDATE(p) ▷ Bottom-up update
    else ▷ If the parent is locked, ...
        $d \leftarrow n.value - v_{\text{past}}$
        for $s \in$ SIBLINGS(n) **do** ▷ ... distribute to siblings
            PROPAGATECHANGE($s, -d$)
        end for
    end if
end procedure

Appendix C

Computation Architecture

Typical Shiny applications depend upon a reactive programming model: A change of an input control in the web browser is sent to R for calculation, and the results are returned to update the visualization. However, each user interaction in *Argus* can potentially trigger a time-consuming computation that could render the user interface unresponsive. *Argus* thus only uses Shiny to provide direct communication between R and JavaScript [Cheng, 2018]. R returns computational results to JavaScript asynchronously, which ensures that the interface remains responsive. Figure C.1 shows a sample scenario: After receiving input A, the simulator computes a preview (Results $A_{1–30}$) and sends it back to JavaScript to be visualized. Subsequent results are sent back to JavaScript until the computation is complete. Suppose the user triggers Input B while the previous simulation is still running. *Argus* calculates the preview results ($B_{1–30}$) and pushes the updates to the user interface. Remaining calculations of parameter set A are calculated in parallel on a separate worker process and gradually sent back to the JavaScript side for storage. When the user revisits an earlier history point, previously stored output immediately shows results without requiring additional computation, which makes the history view fully responsive.

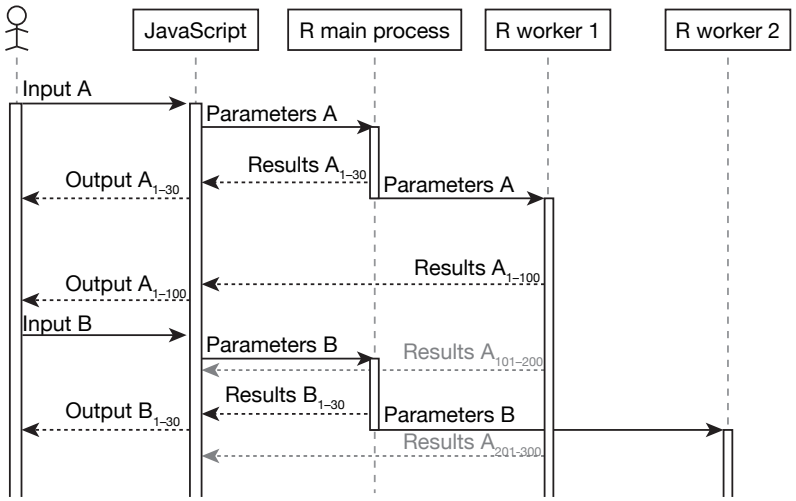


Figure C.1: A sequence diagram shows how *Argus* progressively receives and displays simulation results for a responsive user interface. Grey lines represent the results that are not shown on screen but stored for use when the user navigates back through the history.

Appendix D

Think-aloud study

To validate *Argus*, we conducted an observational study that captures participants' exploration process and insights on power analysis. We focused on the following research question: what insights can researchers gain from being able to interactively explore the impact of design choices (e.g., number of replication, number of participants, counterbalancing) for their experiments.

D.1 Method

We used a think-aloud protocol where participants voice their observations and reasoning [Lewis, 1982], and then performed a qualitative analysis of the results with affinity diagramming. Our analysis focused on *insights* that participants gained [Saraiya et al., 2005]. The study design and the analysis plan are preregistered at [click here for an anonymized URL on osf.io], and were conducted as such—unless stated otherwise below.

D.1.1 Participants

We recruited nine male participants from four different research labs, in four different countries, including junior (Ph.D. candidates) and senior researchers. Note that the types of insights that each participant

might gain from using *Argus* depends on their prior experience with experiment design and *a priori* power analysis, and may not correlate directly with their academic level. However, researchers trained by the same institution may share the same experiment design philosophy. We interviewed each participant about their prior experience with planning, conducting, and analyzing experiment data, and classified them as novice ($P_{i\mathrm{N}}$) or experienced ($P_{i\mathrm{E}}$). For the latter group, researchers have several years of experience with controlled experiments in the field of HCI and/or VIS. Three participants (2 experienced, 1 novice) participated locally and the rest participated remotely. Each participant received the equivalent of a 30 EUR gift card.

	Expertise	Academic Level	Country		
$P1_E$	Experienced	Senior Scientist	FR	Lab1	
$P2_E$	Experienced	Senior Scientist	FR	Lab1	
$P3_N$	Novice	Ph.D. Student	FR	Lab1	
$P4_E$	Experienced	Post-doc	DK	Lab2	(remote)
$P5_E$	Experienced	Ph.D. Student	DE	Lab3	(remote)
$P6_N$	Novice	Ph.D. Student	DK	Lab2	(remote)
$P7_N$	Novice	Ph.D. Student	DK	Lab2	(remote)
$P8_N$	Novice	Ph.D. Student	CH	Lab4	
$P9_N$	Novice	Post-doc	FR	Lab1	(remote)

Table D.1: Background information of the participants.

D.1.2 Apparatus

We used an earlier version of Argus that did not include the whole-experiment practice effect in the *Confound* sliders. Local participants used *Argus* installed on a Macbook Pro (15-inch, 2.5GHz, MacOS 10.14), with QuickTime to record their screens. Remote participants accessed *Argus* via Shinyapps.io, with Skype for their interviews and screen recordings.

D.1.3 Procedure

Training: After the participants gave an informed consent, they watched a video provided in the supplemental material. The video provides a short refresher on experiment design and statistics and gives an overview of *Argus*. During the video, two prompts encourage participants to pause and try out interactions with *Argus*. Participants

can then freely try adjusting the parameters in a dummy experiment setting. Participants are encouraged to ask questions or seek clarifications. Prior to the task, we asked participants to ensure that they were able to use and felt comfortable using *Argus*.

Testing: The participants were asked to determine the sample size for a Fitts's law experiment similar to Douglas et al. [1999]. The experiment compares two devices (a touchpad and a joystick) at three indices of difficulties (3, 5, and 7). To simulate prior domain knowledge for estimating effect sizes, each participant received an information package (pp. 7–11 of the preregistration) printed on paper:

1. A summary of Douglas et al. [1999]'s study with the overall means and SD of the movement time for each device. To simulate the prior knowledge about confounding variables, we also indicate that there was a mild learning effect. To simulate constraints in experiment planning, the description indicated that participants were tired at the end of the original experiment.

2. Excerpts from a trial table, with three Latin-square counterbalancing strategies: (1) Douglas et al. [1999]'s original design: blocked by the device variable; (2) blocked by the device variable and serial-order where all trials with the same device are performed back-to-back; and (3) serial-order by device without blocking.

Before starting to use *Argus*, participants can ask questions about these materials, but may not ask questions during the session.

The main task is to explore the parameters and find a realistic sample size given the resource constraints typical of experiments they have previously conducted. We also ask them to propose two other variants of the experiment design that would reduce the overall number of trials, given the participant fatigue indicated in the information package. We regularly remind participants to verbally describe their actions and to 'think aloud'.

Post-task questionnaire and interview: Participants rate their experience and the insights they gained during the study on a 5-point Likert-style questionnaire. We then interview them about the process and probed for further insights they gained about power analysis.

D.1.4 Data Collection and Analysis

We video-recorded the screen, logged the interaction steps, recorded audio, and took field notes. Two of the co-authors used the field notes to guide a partial transcription for points that the participants voiced, including observations and insights. The transcriptions were coded with a top-down coding scheme based on the typology of data models [Choi et al., 2019]. Three of the co-authors performed a bottom-up thematic analysis together using the affinity diagram method [Holtzblatt and Beyer, 2016].

In addition to the preregistered analysis, we also extract how the users move from clicking on one input control to another and calculated first-order transition probabilities. Although the transition probabilities did not capture how the users attend to views that does not require clicking—e.g., *Pairwise-difference* view and *Power Trade-off* view—they can indicate how the users explore the parameter space.

D.2 Results

This section describes our observations of the participants' interactions with *Argus* and the insights they, and we gained. We will use "users" to refer to the participants in our study to avoid confusion with the "number of participants" term in *Argus*.

Overall, the majority of the users reported that they have gained new insights about experiment design (Figure D.1): *"the preview is very useful to understand the confound effects."*($P9_N$). $P7_N$, $P8_N$ were not familiar with carry-over effect and practice effect but they expressed their understanding of the difference between these effects when they saw the previews. Five users applied their experience in conducting experiment to consider potential confounds. For example, $P8_N$ said *"adding more replications can yield higher power but participants may be tired [so] I need to increase the fatigue."* after increased the number of replications.

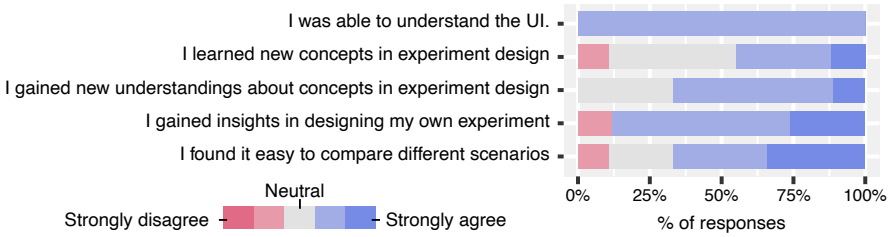


Figure D.1: Result of Study Questionnaire.

D.2.1 Causal Inference about Parameter Relationships

Based on the interview and screen recording video, we coded users' expression of causality (e.g., changing X affect Y) between power analysis parameters. The results is shown in Table D.2. The most frequent insights connect the number of replications and the number of participants to the power (Table D.2 row A and B): *"The power is very high now. I am going to tweak replications and participants to see how power is going to change [...] reduce the number of participants, power drops down. It makes sense"* ($P4_E$). Participants also interpret the characteristics of the curve in *Power Trade-off* view: *"The power get stabled after a certain number of participants. The current number of participant is a bit too much. We can reduce the number"* ($P5_E$). These results are within our expectation because the *Power Trade-off* view directly shows this relationship.

	From	To	Count	Participants
A	# replications	power	6	$P3_N$, $P5_E$, $P6_N$, $P7_N$, $P8_N$, $P9_N$
B	# participants	power	5	$P1_E$, $P3_N$, $P4_E$, $P8_N$, $P9_N$
C	expected means	power	2	$P1_E$, $P2_E$
D	fatigue effect	power	2	$P7_N$, $P9_N$
E	experiment design	power	2	$P4_E$, $P8_N$
F	expected means	conf. interval	2	$P2_E$, $P5_E$
G	experiment design	fatigue effect	2	$P3_N$, $P5_E$
H	# replications	fatigue effect	1	$P6_N$
I	practice effect	power	1	$P8_N$

Table D.2: Causality insights that the participants made, based on the coding of the interview data and screen recording video.

According to the transition probabilities, the users switches between manipulating the group-means and the grand-mean during their exploration (Figure D.2, A). This result demonstrates the usefulness of these controls on top of the normal bar charts. The causal link between the expected mean to power and to the confidence intervals

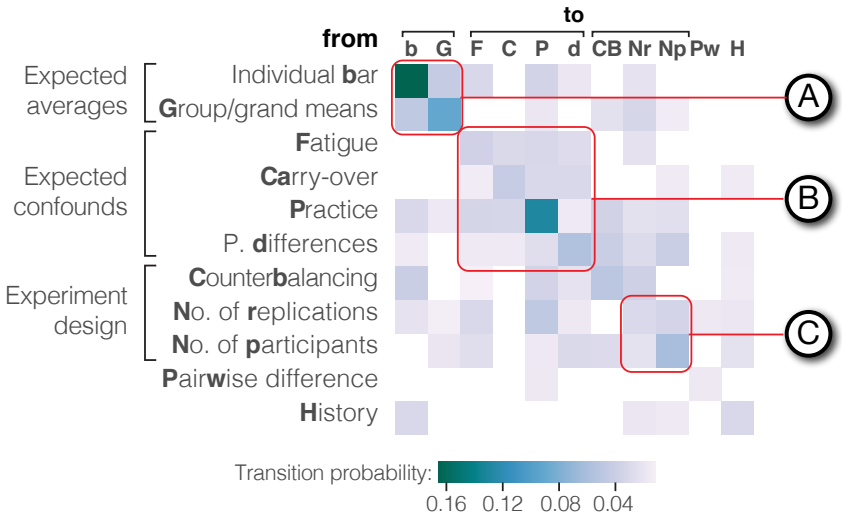


Figure D.2: Average transition probabilities among the input controls, averaged across participants. Three groups of controls (A–C) tends to be more frequently used together than others.

in the *Pairwise-difference* view were expressed by two users each (Table D.2 row C and F). For example, $P1_E$ said *"now I am going to reduce power [. . .] a lot"* after dragging two group-means close to each other.

The confound sliders had frequent transitions among themselves (Figure D.2, B), indicating that confounding effects were explored iteratively together by the users. Even though the carry-over effect was not mentioned on the information sheet, $P9_N$ felt it was necessary to consider it because *"there should be some [carry-over] effect between the first condition and the rests."*

The exceptionally high transition probability from the practice effect slider to itself indicates that the users were more engaged in this effect more than others. This is opposite to Table D.2 (row I) that only $P8_N$ links the practice effect to the power causally. We re-watch the interaction videos and found the reason of this contradiction. The users adjusts the confound sliders with an expectation to see the practice effect's influence. However, because of the initial value of the counterbalancing design (Latin Square & no serialization as used by Douglas et al. [1999]) and the number of replications (1 replication) does not allow the practice effect to manifest. In summary, the results suggests that when the causal link between the parameter and the power

is moderated by the choices of the experiment design parameters, it could be more difficult for the users to make a set of parameters that can demonstrate the connection.

D.2.2 The Use of the *History* view

Five users tweaked expected confounds and observe how the power of adjacent nodes in the *History* view gradually changes. Four users repeatedly used the hover function to preview the difference. Two expert users use the branching to explore multiple strands of parameter configurations. These behaviors show that the *History* view successfully facilitates the exploration of statistical power.

Appendix E

A priori power analysis practices at CHI

Caine [2016] conducted a systematic literature survey on sample sizes at CHI, and found that the median sample size is 18 with 50% of studies reporting fewer than 18 participants. To complement Caine's work, we conducted a literature review to investigate the *a priori* power analysis practices at CHI between 2016 and 2020.

E.1 Method

We used CERMINE [Tkaczyk et al., 2015] to extract the content of all papers and filtered them for "experiment" and "power analys"[1]. For the resulting papers, we manually checked if the authors were using an *a priori* power analysis to plan the sample size for their experiment. Furthermore, we recorded how the actual sample size differed from the planned one. We identified four different categories of experiments: lab, online, crowdsourcing, and non-human samples. The latter category included, for example, online advertisements as samples. A total of six papers did not conduct an *a priori* power analysis, but a post-hoc power analysis. We excluded those results from the analysis.

[1] To include both singular and plural form.

E.2 Results and Discussion

CHI		2015	2016	2017	2018	2019	2020
Full papers using experiment		315	369	391	464	474	507
Used power analysis to plan sample size		1	3	3	8	6	9
Actual sample is ... power analysis sample size.	... higher than ...	0	1	2	2	2	4
	... same as or equal to ...	0	2	1	1	2	2
	... lower than ...	0	0	0	2	0	0
Unstated sample size from power analysis		1	0	0	3	2	3
Settings		L: 1 O: 0 C: 0 N: 0	L: 1 O: 2 C: 0 N: 0	L: 2 O: 0 C: 1 N: 0	L: 2 O: 3 C: 2 N: 1	L: 2 O: 0 C: 3 N: 1	L: 4 O: 0 C: 5 N: 0

L: Lab, **O**: Online, **C**: Crowdsourcing, **N**: Non-human samples

Table E.1: Literature review summary.

Table E.1 shows the proportion of papers with an *a priori* power analysis. It is evident that power analysis seems to be unimportant at CHI for planning sample sizes. We believe that power analysis is complex due to the dynamic relationship between input and output parameters. Wang et al. [2021] created a detail task analysis outlining the challenges of performing such analysis.

Appendix F

Data Analysis and Result Adjustments

First, we explain the calculation of the adjusted p-value with stagewise ordering. Second, we describe the conceptual model and available packages for mean estimate adjustment from which effect sizes such as Cohen's d can be calculated. Lastly, we outline the confidence intervals adjustment.

F.1 Calculation of adjusted p-values with stagewise ordering

In fixed-sample design, a p-value is defined as the probability of obtaining an effect that is at least as extreme as the observed effect assuming that the null-hypothesis is true [Proschan et al., 2006]. This definition needs to be adopted to fit the monitored data collection during SED: The adjusted p-value is defined as the probability of obtaining an effect that is at least as extreme as the observed effect assuming that the null-hypothesis is true **and the previous analyses were not significant**. Several mathematical models can be used for this adjustment, however, *stagewise ordering* is the most common one. *Stagewise ordering* only relies on the preceding interim α boundaries and the observed result while others also take into account the future interim α boundaries [Proschan et al., 2006, Wittes, 2012]. The adjusted p-value

p can be calculate as follows:

$$p = Pr(\bigcup_{i=1}^{j-1} (Z(t_i) \geq c_i) \cup Z(t_j) \geq z_j), \tag{F.1}$$

where $Z(\tau)$ is the resulting interim z-score at information time τ based on the conditional distribution; $c_1, \ldots, c_{j-1}$ are the planning boundaries at analyses $1, \ldots, j-1$; and z_j is the observed z-score at the jth analysis.

For example, let's consider the study planned according to Table 4.4, and assume that the researcher stopped the data collection at the third analysis, i.e. $\tau = 0.75$, with a z-score of 2.7 (p = 0.0069). The calculation for the adjusted p-value using stagewise ordering is:

$$p = Pr(Z(0.25) \geq 4.37 \cup Z(0.5) \geq 2.81 \cup Z(0.75) \geq 2.7) = 0.0119 \tag{F.2}$$

F.2 Calculation of mean differences based on drift θ

A sequential experimental design is considered a Brownian motion process $W(t)$ with linear drift θ. Brownian motion is a continuous-time stochastic process that can be used to model the study's outcome. Furthermore, increments of the process are independent from each other, which resembles recruitment of participants, i.e. participant 1 performs independently of participants 2. The expected value of a Brownian motion process is 0, i.e. $E[W(t)] = 0$. The drift θ describes the rate at which this expected value $E[W(t)]$ changes over time. In sequential experimental design, the drift θ can be understood as the z-score that is expected at the end of the study. In order to calculate the mean difference and its confidence interval, we start by estimating the drift parameter θ. The expected value for the observed drift parameter $E(\theta)$[1] can be expressed as:

$$E(\theta) = \sum_{i=1}^{M} E\left\{ \frac{W(T)}{T} \middle| T = t_i \right\} Pr(T = t_i), \tag{F.3}$$

where M is the number of analyses, and $\frac{W(T)}{T}$ is the maximum likelihood estimate of θ given that $T = t_i$. The equations for the drift numerical calculation would exceed the scope of this paper, but can be

[1] To make the formulas simpler to parse, we do not differentiate between the true drift θ, and the observed or estimated drift $\hat{\theta}$.

found with examples in [Li and DeMets, 1999, Proschan et al., 2006, Jennison, 1999]. Luckily, there are R packages such as `gsDesign` [Anderson, 2020] and `GroupSeq` [Pahl, 2018] that can perform this computation effortlessly, e.g., `findDrift(...)`.

Once the drift parameter is known, the mean difference M_{diff} can be calculated as follows:

$$M_{diff} = \theta \times \sqrt{\frac{2\sigma^2}{N}}, \tag{F.4}$$

where θ is the drift parameter, σ is the pooled standard deviation (using the planned final sample size even if the study was stopped early), and N is the planned final sample size.

Appendix G

Simulations

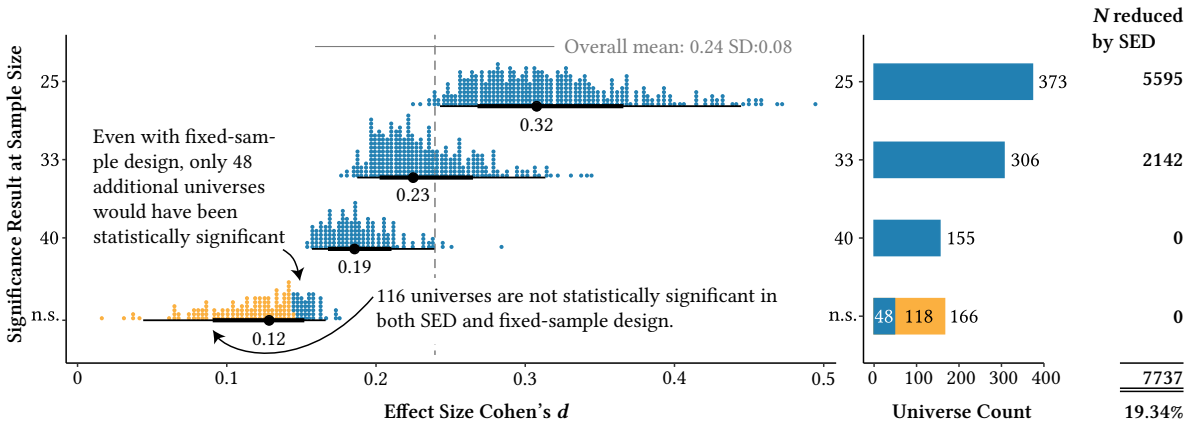


Table G.1: The distribution of Cohen's d from a simulation of 1,000 universes based on the demonstration study. The results are juxtaposed along the vertical axis according to whether and when the results are statistically significant with SED plan. In the n.s. row, 48 universes would have yielded a statistically significant result with the fixed-sample design used. Right: frequency and saving summary.

We ran two simulations to assess the benefits of SED in the long run for lab studies and online crowdsourcing studies. The reproducible R code used in this section is provided in Supplementary Material S2.

The first simulation represents lab studies. We use the same setup as in section G and simulate running the same study in 1,000 different universes. Table G.1 presents the effect sizes in a quantile dot plot [Kay et al., 2016a, Kay, 2020]. With SED, 83.4% of the universes found that the effect was statistically significant. With a fixed-sample design

instead of SED, only an additional 4.8% are statistically significant, all of which yield an effect size lower than 0.2— Cohen's "small" criterion [Cohen, 1988]. Moreover, regardless of design method, the results are not statistically significant in 11.8% of the cases. On the other hand, SED reduced the number of participants by 19% (7,737 participants) in total across all universes, demonstrating the potential of SED to save significant resources.

Hypotheses: There are four null hypothesis on the same dependent variable. SESOI ±0.06

	N	nominal α
Interim Analysis 1	1200	.0015
Interim Analysis 2	1800	.0147
Final Analysis	2351	.0338

SED decision rules: Stop when either:
• all p-values < nominal α → statistically significant
• all Cohen's d < SESOI → futility

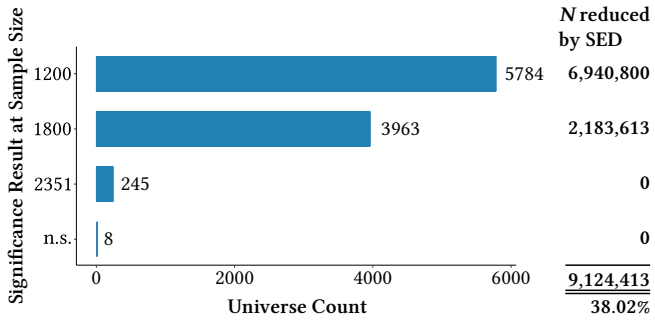


Table G.2: A simulation of a crowdsourcing study [Hofman et al., 2020] in 1,000 universes with the same SED plan shown on the left. The right chart shows that most studies could have stopped early and only a small number of universes did not show significant results.

The second simulation represents crowdsourcing studies. In these studies, increasing the number of participants is relatively easier than in lab studies. However, some data points need to be removed because participants failed attention checks, generated extreme outliers, or faced technical problems [Komarov et al., 2013]. Therefore, crowdsourcing studies tend to recruit more participants than required by *a priori* power analysis—see Appendix E for statistics on this practice in CHI papers.

We ran a simulation using the data from [Hofman et al., 2020] with all of their four hypotheses. To stop early, all hypotheses need to have their p-values below the nominal α. The SED plan and the results are shown in Table G.2. Assuming the same order of participants as the original data, this study could have stopped after 1,200 participants. Since the original experiment had 2,351 participants, SED would have reduced the cost by (2,351 − 1,200) × 0.75 USD = 863.25 USD. Therefore, even when the cost per participant is low, SED can be substantially economical.

Bibliography

Jacob Abbott, Haley MacLeod, Novia Nurain, Gustave Ekobe, and Sameer Patil. Local standards for anonymization practices in health, wellness, accessibility, and aging research at chi. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–14, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300692. URL https://doi.org/10.1145/3290605.3300692.

Keaven Anderson. *gsDesign: Group Sequential Design*, 2020. URL https://CRAN.R-project.org/package=gsDesign. R package version 3.1.1.

Keaven M. Anderson and Jason B. Clark. Fitting spending functions. *Statistics in Medicine*, 29(3):321–327, 2009. doi: 10.1002/sim.3737. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3737.

Samantha F. Anderson and Ken Kelley. *BUCSS: Bias and Uncertainty Corrected Sample Size*, 2019. URL https://CRAN.R-project.org/package=BUCSS. R package version 1.1.0.

Samantha F. Anderson, Ken Kelley, and Scott E. Maxwell. Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, 28(11):1547–1562, 2017. doi: 10.1177/0956797617723724. URL https://doi.org/10.1177/0956797617723724. PMID: 28902575.

Peter Armitage, CK McPherson, and BC Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, 132(2):235–244, 1969.

Ignacio Avellino, Sheida Nozari, Geoffroy Canlorbe, and Yvonne Jansen. Surgical video summarization: Multifarious uses, sum-

marization process and ad-hoc coordination. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021. doi: 10.1145/3449214. URL https://doi.org/10.1145/3449214.

Thom Baguley. Understanding statistical power in the context of applied research. *Applied Ergonomics*, 35(2):73–80, 2004. doi: 10.1016/j.apergo.2004.01.002.

Monya Baker. 1500 scientists lift the lid on reproducibility. *Nature*, 533 (11):452–454, 2016. doi: 10.1038/533452a.

Arthur Bakker, Jinfa Cai, Lyn English, Gabriele Kaiser, Vilma Mesa, and Wim Van Dooren. Beyond small, medium, or large: points of consideration when interpreting effect sizes. *Educational Studies in Mathematics*, 102(1):1–8, 2019.

Louise Barkhuus and Jennifer A Rode. From mice to men-24 years of evaluation in chi. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, volume 10, 2007.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

P Bauer. Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie*, 20(4):130–148, 1989.

Peter Bauer, Frank Bretz, Vladimir Dragalin, Franz König, and Gernot Wassmer. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine*, 35(3):325–347, 2016. doi: 10.1002/sim.6472. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6472.

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, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan,

Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, Jan 2018. ISSN 2397-3374. doi: 10.1038/s41562-017-0189-z. URL https://doi.org/10.1038/s41562-017-0189-z.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

A. F. Blackwell, C. Britton, A. Cox, T. R. G. Green, C. Gurr, G. Kadoda, M. S. Kutar, M. Loomes, C. L. Nehaniv, M. Petre, C. Roast, C. Roe, A. Wong, and R. M. Young. Cognitive dimensions of notations: Design tools for cognitive technology. In Meurig Beynon, Chrystopher L. Nehaniv, and Kerstin Dautenhahn, editors, *Cognitive Technology: Instruments of Mind*, pages 325–341, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44617-0.

Michael Borenstein, Jacob Cohen, Hannah R. Rothstein, Simcha Pollack, and John M. Kane. A visual approach to statistical power analysis on the microcomputer. *Behavior Research Methods, Instruments, & Computers*, 24(4):565–572, 1992. doi: 10.3758/BF03203606. URL https://doi.org/10.3758/BF03203606.

G. E. P. Box and K. B. Wilson. *On the Experimental Attainment of Optimum Conditions*, pages 270–310. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_23. URL https://doi.org/10.1007/978-1-4612-4380-9_23.

Andrew Brand, Michael T. Bradley, Lisa A. Best, and George Stoica. Accuracy of effect size estimates from published psychological research. *Perceptual and Motor Skills*, 106(2):645–649, 2008. doi: 10.2466/pms.106.2.645-649. URL https://doi.org/10.2466/pms.106.2.645-649.

Andrew Brand, M. T. Bradley, Lisa A. Best, and George Stoica. Accuracy of effect size estimates from published psychological experiments involving multiple trials. *The Journal of General Psychology*, 138(4):281–291, 2011. doi: 10.1080/00221309.2011.

604365. URL https://doi.org/10.1080/00221309.2011.604365. PMID: 24836566.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa. URL https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

Matthew Brehmer and Tamara Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013. doi: 10.1109/TVCG.2013.124.

Sebastian Bremm, Tatiana von Landesberger, Juergen Bernard, and Tobias Schreck. Assisted descriptor selection based on visual comparative data analysis. *Computer Graphics Forum*, 30(3):891–900, 2011. doi: 10.1111/j.1467-8659.2011.01938.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2011.01938.x.

Frank Bretz, Franz Koenig, Werner Brannath, Ekkehard Glimm, and Martin Posch. Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*, 28(8):1181–1217, 2009. doi: 10.1002/sim.3538.

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376, 2013. doi: 10.1038/nrn3475. URL https://doi.org/10.1038/nrn3475.

Kelly Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 981–992, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858498. URL https://doi.org/10.1145/2858036.2858498.

Paul Cairns. *Doing better statistics in human-computer interaction*. Cambridge University Press, 2019.

Gregory Campbell. Similarities and differences of bayesian designs and adaptive designs for medical devices: A regulatory view. *Statistics in Biopharmaceutical Research*, 5(4):356–368, 2013. doi: 10.1080/19466315.2013.846873. URL https://doi.org/10.1080/19466315.2013.846873.

Michael J. Campbell. *Statistics at Square One* -. John Wiley & Sons, New York, 2021. ISBN 978-1-119-40130-8.

Evan C. Carter and Michael E. McCullough. Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, 5, 2014. ISSN 1664-1078. doi: 10.3389/fpsyg.2014.00823. URL https://www.frontiersin.org/article/10.3389/fpsyg.2014.00823.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. ISSN 08834237. URL http://www.jstor.org/stable/2246015.

Stephane Champely. *pwr: Basic Functions for Power Analysis*, 2018. URL https://CRAN.R-project.org/package=pwr. R package version 1.2-2.

Stephane Champely, Claus Ekstrom, Peter Dalgaard, Jeffrey Gill, Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic, Helios De Rosario, and Maintainer Helios De Rosario. *Package 'pwr'*, 2018. URL https://CRAN.R-project.org/package=pwr. R package version 1.2.2.

Shi-Yi Chen, Zhe Feng, and Xiaolian Yi. A general introduction to adjustment for multiple comparisons. *Journal of thoracic disease*, 9 (6):1725–1729, 06 2017. doi: 10.21037/jtd.2017.05.34. URL https://pubmed.ncbi.nlm.nih.gov/28740688.

Joe Cheng. Communicating with shiny via javascript. https://shiny.rstudio.com/articles/communicating-with-js.html, May 2018.

In Kwon Choi, Taylor Childers, Nirmal Kumar Raveendranath, Swati Mishra, Kyle Harris, and Khairi Reda. Concept-driven visual analytics: An exploratory study of model- and hypothesis-based reasoning with visualizations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 68:1–68:14, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300298. URL http://doi.acm.org/10.1145/3290605.3300298.

William S. Cleveland and Robert McGill. An experiment in graphical perception. *International Journal of Man-Machine Studies*, 25 (5):491–500, 1986. doi: https://doi.org/10.1016/S0020-7373(86)80019-0. URL http://www.sciencedirect.com/science/article/pii/S0020737386800190.

Andy Cockburn, Carl Gutwin, and Alan Dix. Hark no more: On the preregistration of chi experiments. In *Proceedings of the 2018 CHI*

Conference on Human Factors in Computing Systems, CHI '18, pages 1–12, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173715. URL https://doi.org/10.1145/3173574.3173715.

Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. Threats of a replication crisis in empirical computer science. *Commun. ACM*, 63(8):70–79, July 2020. ISSN 0001-0782. doi: 10.1145/3360311. URL https://doi.org/10.1145/3360311.

Jacob Cohen. The t Test for Means. In *Statistical Power Analysis for the Behavioral Sciences*, pages 19–74. Academic Press, revised ed edition, 1977. doi: 10.1016/B978-0-12-179060-8.50007-4. URL doi.org/10.1016/B978-0-12-179060-8.50007-4.

Jacob Cohen. *Statistical power analysis for the behavioral sciences. 2nd*. Hillsdale, NJ: erlbaum, 1988.

M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, 2014.

Michael Correll, Dominik Moritz, and Jeffrey Heer. Value-suppressing uncertainty palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–11, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174216. URL https://doi.org/10.1145/3173574.3174216.

David Roxbee Cox and Nancy Reid. *The theory of the design of experiments*. CRC Press, 2000.

Geoff Cumming. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. Routledge/Taylor & Francis Group, New York, NY, US, 2012. ISBN 978-0-415-87968-2 (Paperback); 978-0-415-87967-5 (Hardcover).

Geoff Cumming and Robert Calin-Jageman. *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge, 2017.

Peter Cummings. Arguments for and Against Standardized Mean Differences (Effect Sizes). *Archives of Pediatrics & Adolescent Medicine*, 165(7):592, jul 2011. doi: 10.1001/archpediatrics.2011.97. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{&}db=PubMed{&}dopt=

Citation{&}list{_}uids=21727271http://archpedi.jamanetwork.com/article.aspx?doi=10.1001/archpediatrics.2011.97.

Sarah A. Douglas, Arthur E. Kirkpatrick, and I. Scott MacKenzie. Testing pointing device performance and user assessment with the iso 9241, part 9 standard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 215–222, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 0201485591. doi: 10.1145/302979.303042. URL https://doi.org/10.1145/302979.303042.

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

Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. Increasing the transparency of research papers with explorable multiverse analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–15, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300295. URL https://doi.org/10.1145/3290605.3300295.

William D. Dupont. Sequential stopping rules and sequentially adjusted p values: Does one require the other? *Controlled Clinical Trials*, 4(1):3 – 10, 1983. ISSN 0197-2456. doi: https://doi.org/10.1016/S0197-2456(83)80003-8. URL http://www.sciencedirect.com/science/article/pii/S0197245683800038.

Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. *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, pages 1–11, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300447. URL https://doi.org/10.1145/3290605.3300447.

Edgar Erdfelder, Franz Faul, and Axel Buchner. Gpower: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, 28(1):1–11, 1996. doi: 10.3758/BF03203630. URL https://doi.org/10.3758/BF03203630.

Daniele Fanelli and John P. A. Ioannidis. Us studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences*, 110(37):15031–15036, 2013. doi: 10.1073/

pnas.1302997110. URL https://www.pnas.org/doi/abs/10.1073/pnas.1302997110.

Franz Faul and Edgar Erdfelder. Gpower: A priori, post-hoc, and compromise power analyses for ms-dos [computer program]. 2004.

Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, May 2007. ISSN 1554-3528. doi: 10.3758/BF03193146. URL https://doi.org/10.3758/BF03193146.

P.I. Feder, C.T. Olson, D.W. Hobson, M.C. Matthews, and R.L. Joiner. Stagewise, group sequential experimental designs for quantal responses. one-sample and two-sample comparisons. *Neuroscience & Biobehavioral Reviews*, 15(1):129–133, 1991. ISSN 0149-7634. doi: https://doi.org/10.1016/S0149-7634(05)80104-6. URL https://www.sciencedirect.com/science/article/pii/S0149763405801046. A Review of Animal-to-Human Extrapolation: Issues and Opportunities.

Ronald Aylmer Fisher. *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937.

Food and Drug Administration. Adaptive design clinical trials for drugs and biologics. guidance for industry. Technical Report FDA-2018-D-3124, 2019.

Deborah Fry, Kerri Wazny, and Niall Anderson. Using r for repeated and time-series observations. In Judy Robertson and Maurits Kaptein, editors, *Modern Statistical Methods for HCI*, pages 111–133. Springer, 2016.

Brenda Gaydos, Keaven M. Anderson, Donald Berry, Nancy Burnham, Christy Chuang-Stein, Jennifer Dudinak, Parvin Fardipour, Paul Gallo, Sam Givens, Roger Lewis, Jeff Maca, José Pinheiro, Yili Pritchett, and Michael Krams. Good practices for adaptive clinical trials in pharmaceutical product development. *Drug Information Journal*, 43(5):539–556, 2009. doi: 10.1177/009286150904300503. URL https://doi.org/10.1177/009286150904300503.

Andrew Gelman and John Carlin. Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651, 2014.

Daniel G. Goldstein and David Rothschild. Lay understanding of probability distributions. *Judgment and Decision Making*, 9(1):1–14, 2014. ISSN 1930-2975(Print).

Richard Goldstein. Power and sample size via ms/pc-dos computers. *The American Statistician*, 43(4):253–260, 1989. doi: 10.1080/00031305.1989.10475670. URL https://amstat.tandfonline.com/doi/abs/10.1080/00031305.1989.10475670.

Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12, 2016. doi: 10.1126/scitranslmed.aaf5027. URL https://www.science.org/doi/abs/10.1126/scitranslmed.aaf5027.

Julien Gori, Han L. Han, and Michel Beaudouin-Lafon. *FileWeaver: Flexible File Management with Automatic Dependency Tracking*, pages 22–34. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450375146. URL https://doi.org/10.1145/3379337.3415830.

C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):521–530, 2017.

Thomas Green and Alan Blackwell. Cognitive dimensions of information artefacts: a tutorial. In *BCS HCI Conference*, volume 98, pages 1–75, 1998.

T.R.G. Green and M. Petre. Usability analysis of visual programming environments: A 'cognitive dimensions' framework. *Journal of Visual Languages & Computing*, 7(2):131–174, 1996. ISSN 1045-926X. doi: https://doi.org/10.1006/jvlc.1996.0009. URL https://www.sciencedirect.com/science/article/pii/S1045926X96900099.

Saul Greenberg and Bill Buxton. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 111–120, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580111. doi: 10.1145/1357054.1357074. URL https://doi.org/10.1145/1357054.1357074.

Tovi Grossman and Ravin Balakrishnan. The bubble cursor: Enhancing target acquisition by dynamic resizing of the cursor's activation area. In *Proc. Human Factors in Computing Systems*, CHI '05,

pages 281–290, New York, NY, USA, 2005. ACM. ISBN 1-58113-998-5. doi: 10.1145/1054972.1055012. URL http://doi.acm.org/10.1145/1054972.1055012.

Ken Gu, Eunice Jun, and Tim Althoff. Understanding and supporting debugging workflows in multiverse analysis, 2023.

Niklas Hack, Werner Brannath, and Matthias Brueckner. *AGSDest: Estimation in Adaptive Group Sequential Trials*, 2022. URL https://CRAN.R-project.org/package=AGSDest. R package version 2.3.4.

Branimir K Hackenberger. Bayes or not bayes, is this the question? *Croatian medical journal*, 60(1):50–52, 02 2019. doi: 10.3325/cmj.2019.60.50. URL https://pubmed.ncbi.nlm.nih.gov/30825279.

Jake M. Hofman, Daniel G. Goldstein, and Jessica Hullman. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376454. URL https://doi.org/10.1145/3313831.3376454.

F. Hohman, A. Srinivasan, and S. M. Drucker. Telegam: Combining visualization and verbalization for interpretable machine learning. In *2019 IEEE Visualization Conference (VIS)*, pages 151–155, 2019.

Karen Holtzblatt and Hugh Beyer. *Contextual design: Design for life*. Morgan Kaufmann, 2016.

Kasper Hornbæk. Some whys and hows of experiments in human–computer interaction. *Found. Trends Hum.-Comput. Interact.*, 5(4): 299–373, June 2013. ISSN 1551-3955. doi: 10.1561/1100000043. URL http://dx.doi.org/10.1561/1100000043.

Kasper Hornbæk and Effie Lai-Chong Law. Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 617–626, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-593-9. doi: 10.1145/1240624.1240722. URL http://doi.acm.org/10.1145/1240624.1240722.

Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. Is once enough?: On the extent and content of replications in human-computer interaction. In *Proceedings*

of the 32Nd Annual ACM Conference on Human Factors in Computing Systems, CHI '14, pages 3523–3532, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1. doi: 10.1145/2556288.2557004. URL http://doi.acm.org/10.1145/2556288.2557004.

Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.

J. Hullman, M. Kay, Y. Kim, and S. Shrestha. Imagining replications: Graphical prediction discrete visualizations improve recall estimation of effect uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):446–456, Jan 2018. doi: 10.1109/TVCG.2017.2743898.

Irving K. Hwang, Weichung J. Shih, and John S. De Cani. Group sequential designs using a family of type i error probability spending functions. *Statistics in Medicine*, 9(12):1439–1445, 1990. doi: 10.1002/sim.4780091207. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780091207.

Wonil Hwang and Gavriel Salvendy. Number of people required for usability evaluation: The 10±2 rule. *Commun. ACM*, 53(5):130–133, May 2010. ISSN 0001-0782. doi: 10.1145/1735223.1735255. URL http://doi.acm.org/10.1145/1735223.1735255.

John P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2(8):null, 08 2005. doi: 10.1371/journal.pmed.0020124. URL https://doi.org/10.1371/journal.pmed.0020124.

Matthew A Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.

W. Javed and N. Elmqvist. Exploring the design space of composite visualization. In *2012 IEEE Pacific Visualization Symposium*, pages 1–8, 2012.

Christopher Jennison. *Group sequential methods with applications to clinical trials*. Chapman & Hall etc., 1999. ISBN 0-8493-0316-8.

Leslie K. John, George Loewenstein, and Drazen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532, 2012. doi: 10.1177/0956797611430953. URL https://doi.org/10.1177/0956797611430953. PMID: 22508865.

Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy*, 98(6):1325–1348, 1990. doi: 10.1086/261737. URL https://doi.org/10.1086/261737.

Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, 5(1):193–206, March 1991. doi: 10.1257/jep.5.1.193. URL http://www.aeaweb.org/articles?id=10.1257/jep.5.1.193.

Vigdis By Kampenes, Tore Dybå, Jo E. Hannay, and Dag I.K. Sjøberg. A systematic review of effect size in software engineering experiments. *Information and Software Technology*, 49(11):1073 – 1086, 2007. ISSN 0950-5849. doi: https://doi.org/10.1016/j.infsof.2007.02.015. URL http://www.sciencedirect.com/science/article/pii/S0950584907000195.

Matthew Kay. *ggdist: Visualizations of Distributions and Uncertainty*, 2020. URL http://mjskay.github.io/ggdist/. R package version 2.2.0.

Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 5092–5103, New York, NY, USA, 2016a. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858558. URL https://doi.org/10.1145/2858036.2858558.

Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of hci. In *Proc. Human Factors in Computing Systems*, CHI '16, pages 4521–4532, New York, USA, 2016b. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858465. URL http://doi.acm.org/10.1145/2858036.2858465.

Mohamed Khamis, Ludwig Trotter, Ville Mäkelä, Emanuel von Zezschwitz, Jens Le, Andreas Bulling, and Florian Alt. Cueauth: Comparing touch, mid-air gestures, and gaze for cue-based authentication on situated displays. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(4), dec 2018. doi: 10.1145/3287052. URL https://doi.org/10.1145/3287052.

Kyungmann Kim and David L. DeMets. Design and analysis of group sequential tests based on the type i error spending rate function.

Biometrika, 74(1):149–154, 03 1987. ISSN 0006-3444. doi: 10.1093/biomet/74.1.149. URL https://doi.org/10.1093/biomet/74.1.149.

Ross D. King et al. The automation of science. *Science*, 324(5923): 85–89, 2009. ISSN 0036-8075. doi: 10.1126/science.1165620. URL http://science.sciencemag.org/content/324/5923/85.

Roger E. Kirk. Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56(5):746–759, 1996. doi: 10.1177/0013164496056005002. URL https://doi.org/10.1177/0013164496056005002.

Janin Koch, Nicolas Taffin, Andrés Lucero, and Wendy E. Mackay. *SemanticCollage: Enriching Digital Mood Board Design with Semantic Labels*, pages 407–418. Association for Computing Machinery, New York, NY, USA, 2020. ISBN 9781450369749. URL https://doi.org/10.1145/3357236.3395494.

Steven Komarov, Katharina Reinecke, and Krzysztof Z. Gajos. Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 207–216, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990. doi: 10.1145/2470654.2470684. URL https://doi.org/10.1145/2470654.2470684.

R. Kosara and S. Haroz. Skipping the replication crisis in visualization: Threats to study validity and how to address them : Position paper. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pages 102–107, Oct 2018. doi: 10.1109/BELIV.2018.8634392.

John K. Kruschke and Torrin M. Liddell. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, 25(1): 178–206, 2018. doi: 10.3758/s13423-016-1221-4. URL https://doi.org/10.3758/s13423-016-1221-4.

Tze Leung Lai, Philip William Lavori, and Mei-Chiung Shih. Adaptive trial designs. *Annual Review of Pharmacology and Toxicology*, 52(1):101–110, 2012. doi: 10.1146/annurev-pharmtox-010611-134504. URL https://doi.org/10.1146/annurev-pharmtox-010611-134504. PMID: 21838549.

Daniël Lakens. Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44 (7):701–710, 2014a. doi: 10.1002/ejsp.2023. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2023.

Daniël Lakens. Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44 (7):701–710, 2014b. doi: 10.1002/ejsp.2023. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.2023.

Daniel Lakens. The practical alternative to the p-value is the correctly used p-value, Apr 2019. URL psyarxiv.com/shm8v.

Daniël Lakens. Sample Size Justification. *Collabra: Psychology*, 8(1), 03 2022. ISSN 2474-7394. doi: 10.1525/collabra.33267. URL https://doi.org/10.1525/collabra.33267. 33267.

Daniël Lakens and Ellen R. K. Evers. Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, 9(3):278–292, 2014. doi: 10.1177/1745691614528520. URL https://doi.org/10.1177/1745691614528520. PMID: 26173264.

Daniel Lakens, Federico G. Adolfi, Casper J. Albers, Farid Anvari, Matthew A. J. Apps, Shlomo E. Argamon, Thom Baguley, Raymond B. Becker, Stephen D. Benning, Daniel E. Bradford, Erin M. Buchanan, Aaron R. Caldwell, Ben Van Calster, Rickard Carlsson, Sau-Chin Chen, Bryan Chung, Lincoln J. Colling, Gary S. Collins, Zander Crook, Emily S. Cross, Sameera Daniels, Henrik Danielsson, Lisa DeBruine, Daniel J. Dunleavy, Brian D. Earp, Michele I. Feist, Jason D. Ferrell, James G. Field, Nicholas W. Fox, Amanda Friesen, Caio Gomes, Monica Gonzalez-Marquez, James A. Grange, Andrew P. Grieve, Robert Guggenberger, James Grist, Anne-Laura van Harmelen, Fred Hasselman, Kevin D. Hochard, Mark R. Hoffarth, Nicholas P. Holmes, Michael Ingre, Peder M. Isager, Hanna K. Isotalus, Christer Johansson, Konrad Juszczyk, David A. Kenny, Ahmed A. Khalil, Barbara Konat, Junpeng Lao, Erik Gahner Larsen, Gerine M. A. Lodder, Jiří Lukavský, Christopher R. Madan, David Manheim, Stephen R. Martin, Andrea E. Martin, Deborah G. Mayo, Randy J. McCarthy, Kevin McConway, Colin McFarland, Amanda Q. X. Nio, Gustav Nilsonne, Cilene Lino de Oliveira, Jean-Jacques Orban de Xivry, Sam Parsons, Gerit Pfuhl, Kimberly A. Quinn, John J. Sakon, S. Adil Saribay, Iris K. Schneider, Manojkumar Selvaraju, Zsuzsika Sjoerds, Samuel G. Smith, Tim Smits, Jeffrey R. Spies, Vishnu Sreekumar, Crystal N. Steltenpohl, Neil Sten-

house, Wojciech Świkatkowski, Miguel A. Vadillo, Marcel A. L. M. Van Assen, Matt N. Williams, Samantha E. Williams, Donald R. Williams, Tal Yarkoni, Ignazio Ziano, and Rolf A. Zwaan. Justify your alpha. *Nature Human Behaviour*, 2(3):168–171, Mar 2018. ISSN 2397-3374. doi: 10.1038/s41562-018-0311-x. URL https://doi.org/10.1038/s41562-018-0311-x.

K. K. Gordon Lan and David L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 12 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.3.659. URL https://doi.org/10.1093/biomet/70.3.659.

Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, 2017a.

Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. Chapter 3 - experimental design. In Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser, editors, *Research Methods in Human Computer Interaction (Second Edition)*, pages 45 – 69. Morgan Kaufmann, Boston, second edition edition, 2017b. ISBN 978-0-12-805390-4. doi: https://doi.org/10.1016/B978-0-12-805390-4.00003-0. URL http://www.sciencedirect.com/science/article/pii/B9780128053904000030.

Russel V Lenth. Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3):187–193, 2001. doi: 10.1198/000313001317098149.

Clayton Lewis. Using the "thinking-aloud" method in cognitive interface design. Research Report RC 9265 (#40713), IBM Thomas J. Watson Research Center, Yorktown Heights, NY, February 1982. URL https://domino.research.ibm.com/library/cyberdig.nsf/3addb4b88e7a231f85256b3600727773/2513e349e05372cc852574ec0051eea4.

Zhengqing Li and David L. DeMets. On the bias of estimation of a brownian motion drift following group sequential tests. *Statistica Sinica*, 9(4):923–937, 1999. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24306627.

Mark W Lipsey. *Design sensitivity: Statistical power for experimental research*, volume 19. Sage, 1990.

Mark W Lipsey. Design sensitivity: Statistical power for experimental research. In Leonard Bickman and Debra J Rog, editors, *The SAGE handbook of applied social research methods*, volume 19, chapter 2. Sage, 2 edition, 2009.

Wendy E Mackay. Using video to support interaction design. *DVD Tutorial, CHI*, 2(5), 2002.

Wendy E. Mackay, Caroline Appert, Michel Beaudouin-Lafon, Olivier Chapuis, Yangzhou Du, Jean-Daniel Fekete, and Yves Guiard. Touchstone: Exploratory design of experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 1425–1434, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595935939. doi: 10.1145/1240624.1240840. URL https://doi.org/10.1145/1240624.1240840.

Scott Marek, Brenden Tervo-Clemmens, Finnegan J. Calabro, David F. Montez, Benjamin P. Kay, Alexander S. Hatoum, Meghan Rose Donohue, William Foran, Ryland L. Miller, Timothy J. Hendrickson, Stephen M. Malone, Sridhar Kandala, Eric Feczko, Oscar Miranda-Dominguez, Alice M. Graham, Eric A. Earl, Anders J. Perrone, Michaela Cordova, Olivia Doyle, Lucille A. Moore, Gregory M. Conan, Johnny Uriarte, Kathy Snider, Benjamin J. Lynch, James C. Wilgenbusch, Thomas Pengo, Angela Tam, Jianzhong Chen, Dillan J. Newbold, Annie Zheng, Nicole A. Seider, Andrew N. Van, Athanasia Metoki, Roselyne J. Chauvin, Timothy O. Laumann, Deanna J. Greene, Steven E. Petersen, Hugh Garavan, Wesley K. Thompson, Thomas E. Nichols, B. T. Thomas Yeo, Deanna M. Barch, Beatriz Luna, Damien A. Fair, and Nico U. F. Dosenbach. Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660, 2022. doi: 10.1038/s41586-022-04492-9. URL https://doi.org/10.1038/s41586-022-04492-9.

Florian Mathis, Kami Vaniea, and Mohamed Khamis. *RepliCueAuth: Validating the Use of a Lab-Based Virtual Reality Setup for Evaluating Authentication Systems*. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380966. URL https://doi.org/10.1145/3411764.3445478.

Scott E. Maxwell. The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9(2):147–163, 2004. doi: 10.1037/1082-989X.9.2.147. URL https://doi.org/10.1037/1082-989X.9.2.147.

Scott E Maxwell, Michael Y Lau, and George S Howard. Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, 70(6):487, 2015.

JOSEPH E. MCGRATH. Methodology matters: Doing research in the behavioral and social sciences. In RONALD M. BAECKER, JONATHAN GRUDIN, WILLIAM A.S. BUXTON, and SAUL GREENBERG, editors, *Readings in Human–Computer Interaction*, Interactive Technologies, pages 152–169. Morgan Kaufmann, 1995. ISBN 978-0-08-051574-8. doi: https://doi.org/10.1016/B978-0-08-051574-8.50019-4. URL https://www.sciencedirect.com/science/article/pii/B9780080515748500194.

Andrew M McNutt and Ravi Chugh. Integrated visualization editing via parameterized declarative templates. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445356. URL https://doi.org/10.1145/3411764.3445356.

Cyrus R Mehta, Nitin Patel, Pralay Senchaudhuri, and Anastasios Tsiatis. Exact permutational tests for group sequential clinical trials. *Biometrics*, pages 1042–1053, 1994.

Xiaojun Meng, Pin Sym Foong, Simon Perrault, and Shengdong Zhao. *NexP: A Beginner Friendly Toolkit for Designing and Conducting Controlled Experiments*, pages 132–141. Springer International Publishing, Cham, 2017. ISBN 978-3-319-67687-6. doi: 10.1007/978-3-319-67687-6_10. URL https://doi.org/10.1007/978-3-319-67687-6_10.

Mirjam Moerbeek. Bayesian updating: increasing sample size during the course of a study. *BMC Medical Research Methodology*, 21(1):137, 2021. doi: 10.1186/s12874-021-01334-6. URL https://doi.org/10.1186/s12874-021-01334-6.

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

Hans-Helge Müller and Helmut Schäfer. Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57(3):886–891, 2001. doi: 10.1111/j.0006-341X.2001.00886.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2001.00886.x.

Tamara Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press, 2015.

Kevin R Murphy, Brett Myors, and Allen Wolach. *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Routledge, 2014.

Linda K. Muthén and Bengt O. Muthén. How to use a monte carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(4):599–620, 2002. doi: 10.1207/S15328007SEM0904_8.

Vijayan N. Nair, Bovas Abraham, Jock MacKay, John A. Nelder, George Box, Madhav S. Phadke, Raghu N. Kacker, Jerome Sacks, William J. Welch, Thomas J. Lorenzen, Anne C. Shoemaker, Kwok L. Tsui, James M. Lucas, Shin Taguchi, Raymond H. Myers, G. Geoffrey Vining, and C. F. Jeff Wu. Taguchi's parameter design: A panel discussion. *Technometrics*, 34(2):127–161, 1992. ISSN 00401706. URL http://www.jstor.org/stable/1269231.

J. Neyman. Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society*, 105(4): 292–327, 1942. ISSN 09528385. URL http://www.jstor.org/stable/2980436.

J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 20A(1/2):175–240, 1928. ISSN 00063444. URL http://www.jstor.org/stable/2331945.

Jerzy Neyman. Note on an article by sir ronald fisher. *Journal of the Royal Statistical Society. Series B (Methodological)*, 18(2):288–294, 1956. ISSN 00359246. URL http://www.jstor.org/stable/2983716.

Peter C. O'Brien and Thomas R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556, 1979. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2530245.

Peter C O'Brien and Thomas R Fleming. A paired prentice-wilcoxon test for censored paired data. *Biometrics*, pages 169–180, 1987.

Natalia Obukhova. A meta-analysis of effect sizes of chi typing experiments. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery.

Roman Pahl. *GroupSeq: A GUI-Based Program to Compute Probabilities Regarding Group Sequential Designs*, 2018. URL https://CRAN.R-project.org/package=GroupSeq. R package version 1.3.5.

Philip Pallmann, Alun W. Bedding, Babak Choodari-Oskooei, Munyaradzi Dimairo, Laura Flight, Lisa V. Hampson, Jane Holmes, Adrian P. Mander, Lang'o Odondi, Matthew R. Sydes, Sofía S. Villar, James M. S. Wason, Christopher J. Weir, Graham M. Wheeler, Christina Yap, and Thomas Jaki. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16 (1):29, 2018. doi: 10.1186/s12916-018-1017-7. URL https://doi.org/10.1186/s12916-018-1017-7.

C. Papadopoulos, I. Gutenko, and A. E. Kaufman. Veevvie: Visual explorer for empirical visualization, vr and interaction experiments. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):111–120, Jan 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2467954.

E. S. Pearson. Statistical concepts in the relation to reality. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):204–207, 1955. ISSN 00359246. URL http://www.jstor.org/stable/2983954.

Chanda Phelan, Jessica Hullman, Matthew Kay, and Paul Resnick. Some prior(s) experience necessary: Templates for getting started with bayesian analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300709. URL https://doi.org/10.1145/3290605.3300709.

Stuart J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 08 1977. ISSN 0006-3444. doi: 10.1093/biomet/64.2.191. URL https://doi.org/10.1093/biomet/64.2.191.

Stuart J. Pocock. When (not) to stop a clinical trial for benefit. *JAMA*, 294(17):2228–2230, 11 2005. ISSN 0098-7484. doi: 10.1001/jama.294.17.2228. URL https://doi.org/10.1001/jama.294.17.2228.

Fernando P. Polack, Stephen J. Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L. Perez, Gonzalo Pérez Marc, Edson D. Moreira, Cristiano Zerbini, Ruth Bailey, Kena A. Swanson, Satrajit Roychoudhury, Kenneth Koury, Ping Li, Warren V. Kalina, David Cooper, Robert W. Frenck, Laura L.

Hammitt, Özlem Türeci, Haylene Nell, Axel Schaefer, Serhat Ünal, Dina B. Tresnan, Susan Mather, Philip R. Dormitzer, Uğur Şahin, Kathrin U. Jansen, and William C. Gruber. Safety and efficacy of the bnt162b2 mrna covid-19 vaccine. *New England Journal of Medicine*, 383(27):2603–2615, 2020. doi: 10.1056/NEJMoa2034577. URL https://doi.org/10.1056/NEJMoa2034577. PMID: 33301246.

Michael A. Proschan, K. K. Gordon Lan, and Janet Turk Wittes. *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer New York, New York, NY, 2006. ISBN 978-0-387-44970-8. doi: 10.1007/978-0-387-44970-8_7. URL https://doi.org/10.1007/978-0-387-44970-8_7.

Helen C Purchase. *Experimental human-computer interaction: a practical guide with visual examples*. Cambridge University Press, 2012.

Ramana Rao and Stuart K. Card. 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, pages 318–322, New York, NY, USA, 1994. ACM. ISBN 0-89791-650-6. doi: 10.1145/191666.191776. URL http://doi.acm.org/10.1145/191666.191776.

Judy Robertson and Maurits Kaptein. *Modern Statistical Methods for HCI*. Springer, 2016.

Robert Rosenthal. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638–641, 1979. doi: 10.1037/0033-2909.86.3.638. URL https://doi.org/10.1037/0033-2909.86.3.638.

Martin Oliver Sailer. *crossdes: Construction of Crossover Designs*, 2013. URL https://CRAN.R-project.org/package=crossdes. R package version 1.1.

David Salsburg. *The lady tasting tea: How statistics revolutionized science in the twentieth century*. Macmillan, 2001.

P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):443–456, 2005.

Abhraneel Sarma and Matthew Kay. Prior setting in practice: Strategies and rationales used in choosing prior distributions for bayesian

analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376377. URL https://doi.org/10.1145/3313831.3376377.

Abhraneel Sarma, Alex Kale, Michael J Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing alternative data analyses in r notebooks, Apr 2021. URL osf.io/yfbwm.

SAS Institute Inc. *JMP ®13 Design of experiments guide*. SAS Institute Inc., SAS Institute Inc., Cary, NC, USA, 9 2016.

Henry Scheffe. *The analysis of variance*. John Wiley & Sons, 1959.

Alexander M. Schoemann, Patrick Miller, Sunthud Pornprasertmanit, and Wei Wu. Using monte carlo simulations to determine power and sample size for planned missing designs. *International Journal of Behavioral Development*, 38(5):471–479, 2014. doi: 10.1177/0165025413515169. URL https://doi.org/10.1177/0165025413515169.

Felix D. Schönbrodt, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2): 322–339, 2017. doi: 10.1037/met0000061. URL https://doi.org/10.1037/met0000061.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22 (11):1359–1366, 2011.

S. Smart, K. Wu, and D. A. Szafir. Color crafting: Automating the construction of designer quality color ramps. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1215–1225, 2020.

Larisa N Soldatova and Ross D King. An ontology of scientific experiments. *Journal of The Royal Society Interface*, 3(11):795–803, 2006. ISSN 1742-5689. doi: 10.1098/rsif.2006.0134. URL http://rsif.royalsocietypublishing.org/content/3/11/795.

Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi:

10.1177/1745691616658637. URL https://doi.org/10.1177/1745691616658637. PMID: 27694465.

Douglas J. Taylor and Keith E. Muller. Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics - Theory and Methods*, 25(7):1595–1610, 1996. doi: 10.1080/03610929608831787. URL https://doi.org/10.1080/03610929608831787.

Dominika Tkaczyk, Pawel Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Lukasz Bolikowski. Cermine: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, 18(4):317–335, 2015. ISSN 1433-2833. doi: 10.1007/s10032-015-0249-8. URL http://dx.doi.org/10.1007/s10032-015-0249-8.

S. Todd, A. Whitehead, N. Stallard, and J. Whitehead. Interim analyses and sequential designs in phase iii studies. *British journal of clinical pharmacology*, 51(5):394–399, May 2001. ISSN 0306-5251. doi: 10.1046/j.1365-2125.2001.01382.x. URL https://doi.org/10.1046/j.1365-2125.2001.01382.x. bcp1382[PII].

Transparent Statistics in Human–Computer Interaction Working Group. Transparent Statistics Guidelines, Jun 2019. (Available at https://transparentstats.github.io/guidelines).

Lana M. Trick and Zenon W. Pylyshyn. Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review*, 101(1):80–102, 1994. ISSN 1939-1471(Electronic),0033-295X(Print). doi: 10.1037/0033-295X.101.1.80.

Lisa Tweedie, Robert Spence, Huw Dawkes, and Hus Su. Externalising abstract mathematical models. In *Proc. Human Factors in Computing Systems*, CHI '96, pages 406–ff., New York, NY, USA, 1996. ACM. ISBN 0-89791-777-4. doi: 10.1145/238386.238587. URL http://doi.acm.org/10.1145/238386.238587.

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

Fei Wang and Alan E. Gelfand. A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17(2):193–208, 2002. ISSN 08834237. URL http://www.jstor.org/stable/3182824.

Xiaoyi Wang, Alexander Eiselmayer, Wendy E. Mackay, Kasper Hornbaek, and Chat Wacharamanotham. Argus: Interactive a priori power analysis. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):432–442, 2021. doi: 10.1109/TVCG.2020.3028894.

Gernot Wassmer and Friedrich Pahlke. *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis*, 2022. URL https://CRAN.R-project.org/package=rpact. R package version 3.2.3.

Jelte Wicherts, Coosje Veldkamp, Hilde Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel Van Assen. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7:1832, 2016.

Janet Wittes. Stopping a trial early – and then what? *Clinical Trials*, 9(6):714–720, 2012. doi: 10.1177/1740774512454600. URL https://doi.org/10.1177/1740774512454600. PMID: 22879573.

Jacob O. Wobbrock and Julie A. Kientz. Research contributions in human-computer interaction. *Interactions*, 23(3):38–44, apr 2016. ISSN 1072-5520. doi: 10.1145/2907069. URL https://doi.org/10.1145/2907069.

Daniel Wollschläger. *Grundlagen der Datenanalyse mit R*. Springer Berlin Heidelberg, 2017. doi: 10.1007/978-3-662-53670-4. URL http://dx.doi.org/10.1007/978-3-662-53670-4.

Leslie Wu, Jesse Cirimele, Kristen Leach, Stuart Card, Larry Chu, T. Kyle Harrison, and Scott R. Klemmer. Supporting crisis response with dynamic procedure aids. In *Proceedings of the 2014 Conference on Designing Interactive Systems*, DIS '14, pages 315–324, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329026. doi: 10.1145/2598510.2598565. URL https://doi.org/10.1145/2598510.2598565.

Koji Yatani. *Effect Sizes and Power Analysis in HCI*, pages 87–110. Springer International Publishing, Cham, 2016. ISBN 978-3-319-26633-6. doi: 10.1007/978-3-319-26633-6_5. URL https://doi.org/10.1007/978-3-319-26633-6_5.

Zhiyong Zhang. Monte carlo based statistical power analysis for mediation models: methods and software. *Behavior Research Methods*, 46(4):1184–1198, 2014. doi: 10.3758/s13428-013-0424-0. URL https://doi.org/10.3758/s13428-013-0424-0.

Jonathan Zong, Dhiraj Barnwal, Rupayan Neogy, and Arvind Satyanarayan. Lyra 2: Designing interactive visualizations by demonstration. *IEEE Transactions on Visualization and Computer Graphics*, 27 (2):304–314, 2021. doi: 10.1109/TVCG.2020.3030367.

Curriculum vitae

Personal Details

Alexander Eiselmayer
Date of birth: 23 January 1994

Education

August '18 – February '23: Doctoral program at the University of Zurich, Department of Informatics, People and Computing Lab

September '16 – August '18: Master of Science, University of Zurich, Information Systems

October '12 – June '16: Bachelor of Science, Technical University of Munich, Information Systems

Professional experience

January '16 – August '23: Head of IT and Operations at Prio Partners GmbH, Zurich, Switzerland

April '21 – June '21: User Experience Consultant at Netlight Consulting AG, Zurich, Switzerland

October '15 – December '16: Web developer at JKweb GmbH, Zurich, Switzerland