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

Alexander Eiselmayer[1] **Chat Wacharamanotham**[1] **Michel Beaudouin-Lafon**[2] **Wendy E. Mackay**[2]

[1] University of Zurich
Zurich, Switzerland

[2] Univ. Paris-Sud, CNRS,
Inria, Université Paris-Saclay
Orsay, France

eiselmayer@ifi.uzh.ch, chat@ifi.uzh.ch, mbl@lri.fr, mackay@lri.fr

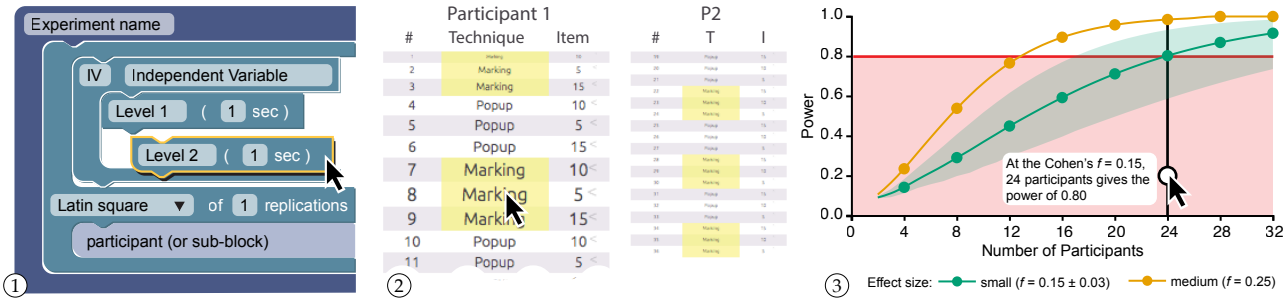

Figure 1. *Touchstone2* experiments consist of interactive "bricks" ① that specify independent variables, blocking, counterbalancing and timing, and generate an interactive trial table ② and an interactive statistical power chart ③.

ABSTRACT

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.

CCS CONCEPTS

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

KEYWORDS

Experiment design; Randomization; Counterbalancing; Power analysis; Reproducibility

ACM Reference Format:
Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. 2019. *Touchstone2*: An Interactive Environment for Exploring Trade-offs in HCI Experiment Design. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3290605.3300447

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." [1]. One explanation is the number of *researcher degrees of freedom*:

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

the methodological decisions from study design up to publication [28], including how many participants are recruited and assigned to which conditions [31]. Cockburn et al. [5] 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 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 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.

Representing experiment designs

Individual research disciplines use various techniques for optimizing experiment designs. For example, industrial manufacturing uses *Response surface design* [2] and the *Taguchi* method [23] 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, Saldatova and King [29] created a computer-readable ontology of scientific experiments (EXPO) that defines terms related to scientific discovery: *research, null* and *alternative hypotheses*, *independent* (IV) and

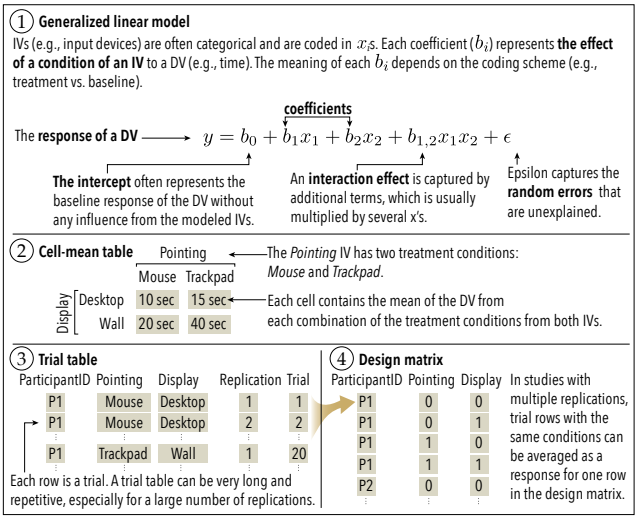

Figure 2. Four experiment designs representations [7][2].

dependent variables (DV), and *results*. This helped automate hypothesis generation and testing for yeast genomics experiments [16]. However, since experiments in this domain are restricted to simple Latin square designs, EXPO omits *blocking* and *counterbalancing*. Papadopoulos et al. [24] present VEEVVIE, an ontology that describes Information Visualization data at the trial level, which unfortunately precludes specifying trial order.

The statistical literature [7, 10] 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 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.

[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.

[2] There are multiple ways to model the error term in a GLM. See dwoll.de/rexrepos/posts/anovaMixed.html based on [32].

[3] *Subjects* is the statistical term; we use *participants* for human subjects.

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 [7]. Some statistical software packages, e.g. *JMP DOE* [27], *Design-Expert*[4], and the R package `skpr` [21] 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 [15], and often include within-participant factors.

The `crossdes` R package [26] generates trial tables and tabulates treatment frequencies by row, column, or concurrence, but only 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* [19] and later *NexP* [20] 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.

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 [11] 30% speed increase, they significantly increase the probability of *misses*, i.e. Type II errors that do not find a real effect (Figure 3).

[4]jmp.com, statease.com

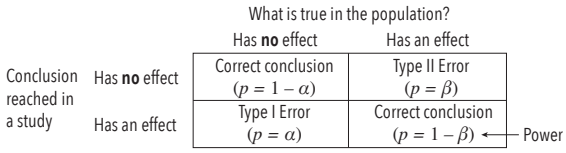

Figure 3. Type I and Type II errors, statistical power.

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` [4], support power analysis. G*Power [9], currently the most comprehensive such, provides a form 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.

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?

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).

Procedure. We interviewed participants at work for 30-60 minutes, using the critical incident technique [18]. 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.

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.

Results

Participants highlighted the following design challenges:

[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

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.

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.

4 DESIGNING *TOUCHSTONE2*

Touchstone introduced a streamlined process for counterbalancing trials [19, Table 1], later adopted by *NexP* [20, Figure 1], with different views accessible in different tabs. The

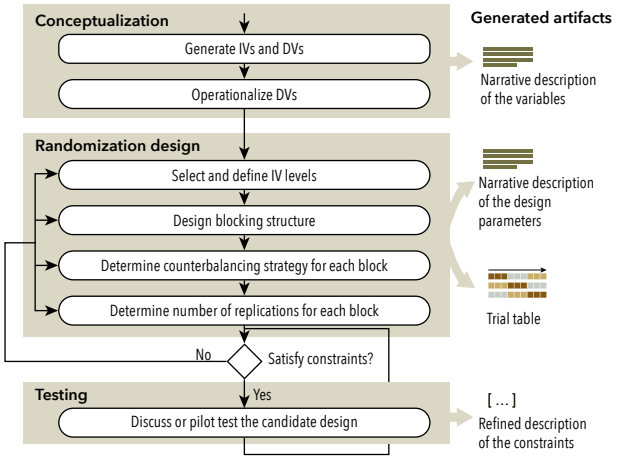

Figure 4. Counterbalancing is highly iterative: Multiple artifacts (right) capture, reveal, and communicate the design.

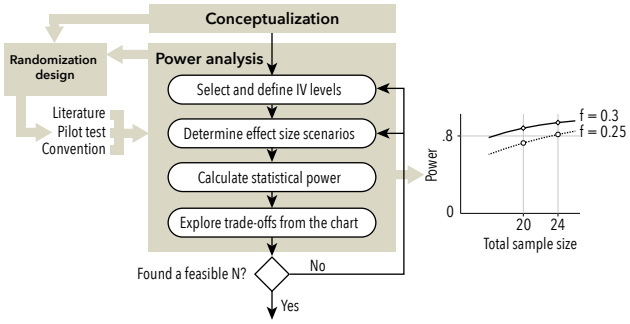

Figure 5. Power chart: Compare several possible effect sizes.

results from our interviews highlight the iterative and collaborative nature of the process, the multiplicity of artifacts generated to communicate designs (Figure 4), and the need to support power analysis (Figure 5).

Counterbalancing process: Researchers generate artifacts (Figure 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.

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 [6, p. 56], and estimating the effect size (Figure 5). The latter is difficult and may discourage users from conducting a power analysis [17, 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?*"[22, p. 17].

To cope with this conundrum, researchers usually visualize the relationships among N, power, and possible effect sizes in a *power chart* (Figure 5, right) to weigh the benefits of more power against the cost of more participants. In Figure 5 (left), increasing the sample size 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.

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.

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

[7]Number of participants

[8]Claiming an effect when one does not exist.

[9]How much DVs (measures) change according to different IV levels.

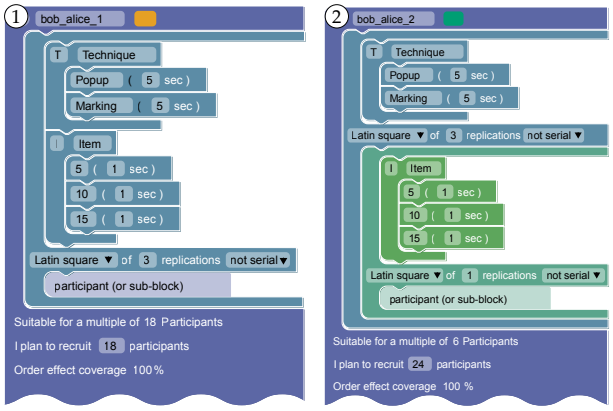

Figure 6. Two blocking strategies for a [2x3] within-participants design to compare POPUP and MARKING menus.

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 6, Design ① is a [2x3] within-participants design to compare menus, where TECHNIQUE has two values: POPUP and MARKING, 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 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 7). As a result, the design is now balanced for multiples of only six participants.

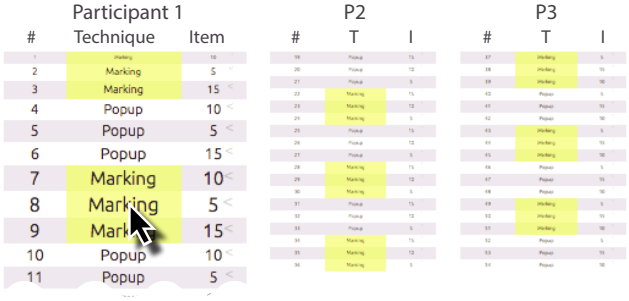

Figure 7. Trial Table Inspection with Fish-eye View

Inspecting trial tables: Manipulating bricks immediately generates a corresponding *trial table* (Figure 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* [30]: 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 [25] 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 7 highlights each MARKING level to show how they are grouped in consecutive trials.

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 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) [13].

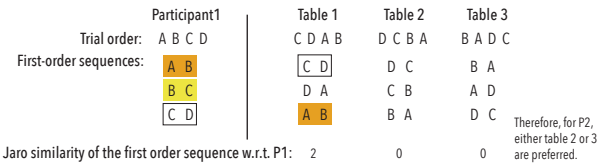

Figure 8. The Jaro similarity measure ensures maximum counterbalancing coverage for each successive participant.

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 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.

[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.

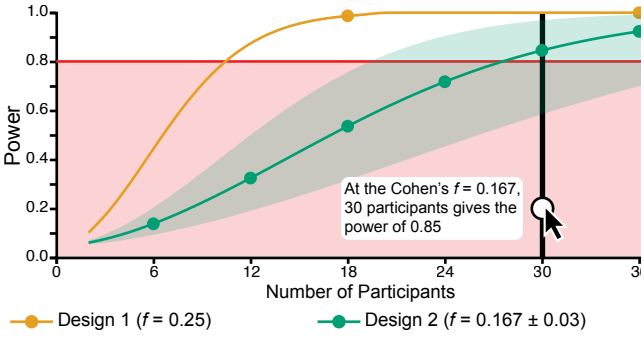

Figure 9. Power analysis: With 18 participants Design 1 is likely to find the effect. Design 2 needs 30 participants.

To refine this estimate, users can choose among Cohen's three conventional effect sizes [6, 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 10).

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.

[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.

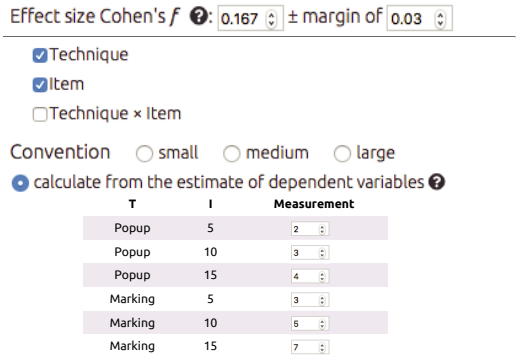

Figure 10. Calculating effect size from pilot data.

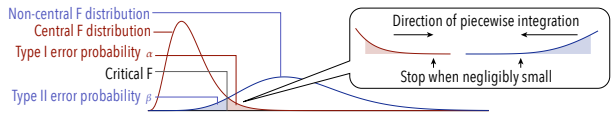

Figure 11. In the power calculation, the direction of integral calculation were optimized for responsiveness.

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 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 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.

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 [19].

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.

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,
                            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.

[12] According to the experiment design and selected effects (Figure 10, top), *Touchstone2* adjusts how the means values (Figure 10, bottom) are aggregated and how the degrees of freedom in the F distributions are calculated from the number of participants. See [9, Table 3] for detailed mathematical formulae.

[13] To produce smooth curves, we calculate power for sample sizes between 1 and 50. ?the sample size of $\{1, 2, \ldots, 50\}$. 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.

[14] https://developers.google.com/blockly/

[15] https://www.w3schools.com/html/html5_webworkers.asp

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

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 preregister HCI experiments [5]. 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.

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.

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 fisheye 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).

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

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.

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*.

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.

Results

We selectively transcribed the audio and video based on field notes. Two authors analyzed the transcripts using thematic analysis [3] 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.

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 5). We also added an explanation about standardized effect sizes and their calculation in the context-sensitive help.

7 DISCUSSION

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

Default parameters and status quo bias

To calculate power, *Touchstone2* uses default parameters and Cohen's conventions [6, 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* [14]. 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.

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 [5], transparent communication of the results [8, 12], and reporting effect sizes [12, 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 [12, 33].) We believe that these aids will persuade researchers to plan experiments with high statistical power instead of *p*-hacking.

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.

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.

ACKNOWLEDGMENTS

This work was partially supported by European Research Council (ERC) grants № 321135 "CREATIV: Creating Co-Adaptive Human-Computer Partnerships" and № 695464 "ONE: Unified Principles of Interaction".

REFERENCES

[1] Monya Baker. 2016. 1500 scientists lift the lid on reproducibility. *Nature* 533, 11 (2016), 452–454. DOI: http://dx.doi.org/10.1038/533452a

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

[2] G. E. P. Box and K. B. Wilson. 1992. *On the Experimental Attainment of Optimum Conditions.* Springer New York, New York, NY, 270–310. DOI: http://dx.doi.org/10.1007/978-1-4612-4380-9_23

[3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. DOI: http://dx.doi.org/10.1191/1478088706qp063oa

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

[5] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK No More: On the Preregistration of CHI Experiments. In *Proc. Human Factors in Computing Systems (CHI '18).* ACM, New York, NY, USA, Article 141, 12 pages. DOI: http://dx.doi.org/10.1145/3173574.3173715

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

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

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

[9] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (01 May 2007), 175–191. DOI: http://dx.doi.org/10.3758/BF03193146

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

[11] Tovi Grossman and Ravin Balakrishnan. 2005. The Bubble Cursor: Enhancing Target Acquisition by Dynamic Resizing of the Cursor's Activation Area. In *Proc. Human Factors in Computing Systems (CHI '05).* ACM, New York, NY, USA, 281–290. DOI: http://dx.doi.org/10.1145/1054972.1055012

[12] Transparent Statistics in Human–Computer Interaction working group. 2018. *Transparent Statistics Guidelines.* Technical Report. DOI: http://dx.doi.org/10.5281/zenodo.1186169 Available at https://transparentstats.github.io/guidelines.

[13] Matthew A Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *J. Amer. Statist. Assoc.* 84, 406 (1989), 414–420.

[14] Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. 1991. Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *Journal of Economic Perspectives* 5, 1 (March 1991), 193–206. DOI: http://dx.doi.org/10.1257/jep.5.1.193

[15] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proc. Human Factors in Computing Systems (CHI '16).* ACM, New York, USA, 4521–4532. DOI: http://dx.doi.org/10.1145/2858036.2858465

[16] Ross D. King and others. 2009. The Automation of Science. *Science* 324, 5923 (2009), 85–89. DOI: http://dx.doi.org/10.1126/science.1165620

[17] Mark W Lipsey. 1990. *Design sensitivity: Statistical power for experimental research.* Vol. 19. Sage.

[18] Wendy E Mackay. 2002. Using video to support interaction design. *DVD Tutorial, CHI* 2, 5 (2002).

[19] Wendy E. Mackay, Caroline Appert, Michel Beaudouin-Lafon, Olivier Chapuis, Yangzhou Du, Jean-Daniel Fekete, and Yves Guiard. 2007. Touchstone: Exploratory Design of Experiments. In *Proc. Human Factors in Computing Systems (CHI '07).* ACM, New York, NY, USA, 1425–1434. DOI: http://dx.doi.org/10.1145/1240624.1240840

[20] Xiaojun Meng, Pin Sym Foong, Simon Perrault, and Shengdong Zhao. 2017. *NexP: A Beginner Friendly Toolkit for Designing and Conducting Controlled Experiments.* Springer International Publishing, Cham, 132–141. DOI: http://dx.doi.org/10.1007/978-3-319-67687-6_10

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

[22] Kevin R Murphy, Brett Myors, and Allen Wolach. 2014. *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests.* Routledge.

[23] 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. 1992. Taguchi's Parameter Design: A Panel Discussion. *Technometrics* 34, 2 (1992), 127–161. http://www.jstor.org/stable/1269231

[24] C. Papadopoulos, I. Gutenko, and A. E. Kaufman. 2016. VEEVVIE: Visual Explorer for Empirical Visualization, VR and Interaction Experiments. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 111–120. DOI: http://dx.doi.org/10.1109/TVCG.2015.2467954

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

[26] Martin Oliver Sailer. 2013. *crossdes: Construction of Crossover Designs.* https://CRAN.R-project.org/package=crossdes R package version 1.1.

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

[28] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22, 11 (2011), 1359–1366.

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

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

[31] Jelte Wicherts, Coosje Veldkamp, Hilde Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel Van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology* 7 (2016), 1832.

[32] Daniel Wollschläger. 2017. *Grundlagen der Datenanalyse mit R.* Springer Berlin Heidelberg. DOI: http://dx.doi.org/10.1007/978-3-662-53670-4

[33] Koji Yatani. 2016. *Effect Sizes and Power Analysis in HCI.* Springer International Publishing, Cham, 87–110. DOI: http://dx.doi.org/10.1007/978-3-319-26633-6_5