

# *Fine-level Control of InfoViz with Foreshadowing via Direct Manipulation*

Master's Thesis  
submitted to the  
Media Computing Group  
Prof. Dr. Jan Borchers  
Computer Science Department  
RWTH Aachen University

*by*  
*Jeanine Marian Bonot*

Thesis advisor:  
Prof. Dr. Jan Borchers

Second examiner:  
Prof. Dr. Ulrik Schroeder

Registration date: 13.08.2018  
Submission date: 13.02.2019



## Eidesstattliche Versicherung

\_\_\_\_\_  
Bonot, Jeanine Marian

\_\_\_\_\_  
362288

Name, Vorname

Matrikelnummer

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Masterarbeit mit dem Titel

### **Fine-level Control of InfoViz with Foreshadowing via Direct Manipulation**

selbständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt. Für den Fall, dass die Arbeit zusätzlich auf einem Datenträger eingereicht wird, erkläre ich, dass die schriftliche und die elektronische Form vollständig übereinstimmen. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

\_\_\_\_\_  
Ort, Datum

\_\_\_\_\_  
Unterschrift

#### **Belehrung:**

##### **§ 156 StGB: Falsche Versicherung an Eides Statt**

Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.

##### **§ 161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt**

(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.

(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.

Die vorstehende Belehrung habe ich zur Kenntnis genommen:

\_\_\_\_\_  
Ort, Datum

\_\_\_\_\_  
Unterschrift



# Contents

<b>Abstract</b>	<b>xv</b>
<b>Acknowledgements</b>	<b>xvii</b>
<b>Conventions</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Statistical Literacy in HCI Research . . . . .	1
1.2 StatPlayground . . . . .	2
1.3 Contribution of This Thesis . . . . .	3
1.4 Research Questions . . . . .	4
1.5 Outline . . . . .	5
<b>2 Related work</b>	<b>7</b>
2.1 Addressing Inadequate Statistical Literacy .	7
2.1.1 StatPlayground: <i>Subramanian and Borchers, 2017</i> . . . . .	8
2.2 Inspiration for Design Techniques . . . . .	8

---

2.2.1	Design by Dragging: <i>Coffey et al., 2013</i>	8
2.2.2	DimpVis: <i>Kondo and Collins, 2014</i> . . .	10
2.2.3	OctoPocus: <i>Bau and Mackay, 2008</i> . . .	10
<b>3</b>	<b>Design Approach</b>	<b>13</b>
3.1	Iterative Design . . . . .	13
3.2	Design Principles . . . . .	14
3.3	Pre-design Decisions . . . . .	14
3.3.1	Representing Summary Statistics . . .	14
3.3.2	Representing Assumption Checks . .	15
3.4	Preliminary User Study . . . . .	16
3.4.1	Fine Control of Properties . . . . .	16
3.4.2	Foreshadowing of InfoViz . . . . .	18
3.4.3	Questionnaire . . . . .	20
	Results . . . . .	21
3.4.4	Participants . . . . .	23
3.5	Abandoned Feature: Simultaneous Manipulation . . . . .	23
3.5.1	Resulting Changes . . . . .	24
<b>4</b>	<b>Interaction Design</b>	<b>25</b>
4.1	Design Layout of StatPlayground . . . . .	25
4.2	Configuring the Dataset . . . . .	26

---

4.3	Manipulating Properties . . . . .	27
4.3.1	Changing Values . . . . .	27
4.3.2	Fine Control of Properties . . . . .	27
4.4	Performing Assumption Checks . . . . .	29
4.5	Viewing the Results . . . . .	31
<b>5</b>	<b>Evaluation</b>	<b>33</b>
5.1	User Study: Phase 1 . . . . .	33
5.1.1	Format . . . . .	33
5.1.2	Tasks . . . . .	34
5.1.3	Participants . . . . .	34
5.1.4	Limitations After Phase 1 . . . . .	35
	Resulting changes . . . . .	36
5.2	User Study: Phase 2 . . . . .	37
5.2.1	Format . . . . .	37
5.2.2	Tasks . . . . .	38
5.2.3	Participants . . . . .	38
5.2.4	Results . . . . .	39
	Quantitative Results . . . . .	39
	Qualitative Results . . . . .	41
5.3	Limitations . . . . .	42
5.4	Discussion . . . . .	43

---

5.4.1	Gestalt Laws . . . . .	43
5.4.2	Shortcomings of Exploration . . . . .	44
<b>6</b>	<b>Summary and Future Work</b>	<b>45</b>
6.1	Summary and Contributions . . . . .	45
6.2	Future Work . . . . .	47
6.2.1	Further Testing . . . . .	47
6.2.2	Inverse Design . . . . .	47
<b>A</b>	<b>Implementation</b>	<b>49</b>
<b>B</b>	<b>User Study</b>	<b>51</b>
B.1	Questionnaire for Experimental Design . . . . .	51
B.1.1	Qualitative Feedback . . . . .	56
	Between-Subjects Design . . . . .	56
	Within-Subjects Design . . . . .	56
B.2	Evaluation . . . . .	57
B.2.1	Scenario Used for Phase 1 of Evaluation	57
B.2.2	Changes to StatPlayground Follow- ing Phase 1 of Evaluation . . . . .	58
	Changes to the Assumption Checks . . . . .	58
	Changes to the Fine Controls . . . . .	58
	Changes to the Results . . . . .	59
	Addition of a Tutorial Page . . . . .	59



Other Changes . . . . .	59
B.2.3 Scenario and Tasks Given in Phase 2 .	60
<b>Bibliography</b>	<b>65</b>
<b>Index</b>	<b>69</b>



# List of Figures

2.1	StatPlayground, 2017 . . . . .	9
2.2	Direct manipulation in Design by Dragging . . . . .	9
2.3	Foreshadowing in DimpVis . . . . .	10
2.4	Foreshadowing in OctoPocus . . . . .	11
3.1	Prototype: Fine-control menu, default . . . . .	16
3.2	Prototype: Fine-control menu, locked . . . . .	17
3.3	Prototype: Fine-control menu, upperbound set . . . . .	18
3.4	Prototype for foreshadowing effect size using vertical lines to indicate interest points . . . . .	19
3.5	Prototype for foreshadowing effect size using a gradient . . . . .	19
3.6	User interpretation of experimental design indicators . . . . .	22
3.7	User design scheme preferences for representing experimental design . . . . .	22
4.1	Screenshot of StatPlayground with two datasets . . . . .	25

---

4.2	Tooltips and cursor styles as signifiers . . . . .	28
4.3	Default fine-controls menu . . . . .	28
4.4	Fine-controls menu with a set upperbound . . . . .	29
4.5	Cursor styles used for the fine-control menu . . . . .	29
4.6	Cursor styles used in StatPlayground . . . . .	30
4.7	Designs for assumption checks: normality . . . . .	30
4.8	Designs for assumption checks: homogeneity of variances . . . . .	31
4.9	Boxplot colours . . . . .	32
4.10	Foreshadowing . . . . .	32
B.1	Scenario used for explaining experimental design . . . . .	52
B.2	Questions for interpretation of experimental design using arrows . . . . .	53
B.3	Questions for interpretation of experimental design using colours . . . . .	54
B.4	Questions asking for the participant's preference for indicating experimental design. . . . .	55
B.5	Paper handout with an overview of terminology used during evaluation . . . . .	61
B.6	Paper handout with example scenario used during evaluation . . . . .	62
B.7	Paper handout with task list for evaluation . . . . .	63

## List of Tables

4.1	Effect sizes and their corresponding Cohen's d value . . . . .	31
4.2	Colours used to show results . . . . .	32
5.1	Accuracy of expected behaviours in the final user study . . . . .	39



# Abstract

Statistical analysis is an important step in a lot of quantitative research, however there is evidence suggesting that there is inadequate statistical literacy in HCI. Many papers in the past years have incorrectly reported or applied statistics in their research, introducing skepticism into their validity. Several approaches have been made to address the problem of inadequate statistical literacy, one of which includes StatPlayground. StatPlayground is a tool that allows user to control different properties of data and observe their effects on the resulting inferential statistics.

We expanded the design of StatPlayground to include more information visualization. We also added fine-control of properties, to allow for more exploration, and foreshadowing of results, to guide and encourage users to explore. In our user study, we found that our design made the learning process of statistics enjoyable. While some features, such as foreshadowing of results and locking of properties, were largely understood by users, other features were shown not to be intuitive. In this thesis, we describe the development of our new features in StatPlayground, summarize our findings, and discuss future work for the project.





# Acknowledgements

I would like to thank my thesis advisor, Prof. Dr. Jan Borchers, and second examiner, Prof. Dr. Ulrik Schroeder for their time and support.

Thank you, Krishna Subramanian, M.Sc., for being my supervisor. I appreciate the guidance, feedback, and advice you have provided throughout my time at the Media Computing Group.

Thanks to all those who donated their time to participate in my user studies. Your feedback is what creates progress in Human-Computer Interaction.

Thank you to all those who have given me moral support throughout my academic journey. Thanks to my parents and siblings in Canada for their unending generosity, understanding, and support. Thanks to my partner, Tobias Stein, and his family for taking care of me. Thanks to my friends, both near and far. Words cannot express how grateful I am for you all.



# Conventions

Throughout this thesis we use the following conventions.

## *Text conventions*

Definitions of technical terms or short excursus are set off in coloured boxes.

**EXCURSUS:**

Excursus are detailed discussions of a particular point in a book, usually in an appendix, or digressions in a written text.

Definition:  
*Excursus*

The whole thesis is written in Canadian English. We use the plural form for the first person. Unidentified third persons are described in singular "they".



# Chapter 1

## Introduction

In this chapter, we describe the problem of statistical illiteracy among researchers in HCI. We discuss its effect on the validation of research and potential causes. Then we introduce StatPlayground, a tool that aims to help fix the problem, and how additional features can help further improve the user experience.

### 1.1 Statistical Literacy in HCI Research

Research is driven by our curiosity to make sense of the world around us. This journey starts by asking questions. Our answers come in the form of jigsaw puzzle pieces, which, individually, do not answer our question until we have put the pieces together to make an image. In the world of research, the puzzle pieces are the data we collect and statistical analysis is what puts the pieces together. While it may not be the most glamorous discipline in HCI, statistical analysis is important if we want to make sense of and validate our findings.

Our ability to navigate through research requires that we have statistical literacy, which expands beyond one's ability to perform statistical analysis. Researchers need to be able to make sense of how findings are reported (e.g., what

Statistics analysis is an important step when answering questions in HCI.

Statistical literacy is a multi-faceted concept.

it means to read “( $M = 4.34, SD = 0.93$ )” or “ $p = .043$ ” in a paper) and need to be able identify potential holes in other works. Statistical literacy is a broad concept, and encompasses not only statistical knowledge, but also literacy skills, mathematical knowledge, context knowledge, and critical thinking [Gal, 2002].

We summarize statistical literacy as the following:

Definition:  
*statistical literacy*

**STATISTICAL LITERACY:**

Statistical literacy is the ability to understand, report, and criticize statistical findings

Evidence of  
inadequate statistical  
literacy in HCI

There is evidence that statistical literacy in HCI is lacking. Cairns [2007] reviewed 41 papers that used some form of inferential statistics. Of the 41 papers, 40 (98%) revealed some sort of error in how statistics was either reported or analyzed. The papers used in Cairns’ study were presented at two conferences by the British Computer Society (BCS HCI ’05, ’06) and two major journals: Human Computer Interaction (HCIJ ’06’) and ACM Transactions on Computer Human Interaction (TOCHI ’06’). This imposes a problem, since papers from such well-esteemed sources are meant to set the standard in HCI literature.

Inadequate statistical  
literacy is a threat to  
HCI research

From his findings, Cairns [2007] expressed concern that “within HCI the standard of statistical analysis is generally quite low for providing convincing results based on NHST.” The lack of statistical literacy is problematic, as the misunderstanding of statistics in researchers can in some cases be grounds for the invalidation of the research itself. The misuse of statistics is not only prevalent in HCI, but in other research areas as well, such as psychology and medicine [Nickerson, 2000, Silva-Ayçaguer et al., 2010]. This is a problem that urgently needs to be addressed.

## 1.2 StatPlayground

Cairns compiled a list of possible approaches to address inadequate statistical literacy in HCI [Cairns, 2007]. One such approach is the introduction of HCI-specific educa-

tion in statistics, which is what StatPlayground [Subramanian and Borchers, 2017] aims to fulfill. StatPlayground is an exploratory tool that aims to address the problem of statistical illiteracy by providing an environment in which users can explore the relationships between statistical properties. It does so by adding the ability to directly manipulate visualizations, via clicking and dragging, and observe how these manipulations affect the results. The goal of StatPlayground is not to replace traditional instruction-based learning, but rather improve certain statistical literacy skills. Such skills include: the ability to make sense of statistical information, identify the relationships between statistical concepts, have better data awareness, and understand statistical procedures better. StatPlayground does not demand that the user has memorized the different formulas for calculating statistics. Calculations are in many cases carried out by specialized software such as IBM SPSS<sup>1</sup> and R<sup>2</sup>. Instead, StatPlayground provides an environment for users to play around with data to understand the relationships between statistical properties in a “learning by doing” approach.

StatPlayground is an interactive tool, which allows user to explore statistical concepts.

### 1.3 Contribution of This Thesis

We added new design features to StatPlayground to further improve learning experience. Our contribution includes the following:

1. **Information visualization (InfoViz).** We aimed to replace text descriptors with symbolic representations. Assumption checks are represented as visualizations to provide a more intuitive meaning for the assumptions they represent. Colours are also used as a way to communicate whether assumptions are satisfied and the varying effect sizes.
2. **Fine-level control.** The previous version of StatPlayground allowed users to use direct manipulation to

Features were added to StatPlayground to limit the amount of text required, to give the user more control, and to encourage exploration.

<sup>1</sup><https://www.ibm.com/analytics/spss-statistics-software>

<sup>2</sup><https://www.r-project.org>

change different properties, such as the mean of a boxplot. We added fine-level control by enabling restriction of properties. These restrictions include locking of properties at particular values and adding upper- and lowerbounds to properties, so that properties are limited to how much they can change.

3. **Foreshadowing of results.** To encourage exploration, we have added a foreshadowing feature, which allows the user to preview the different interest points manipulating a property will cause the results to change.

## 1.4 Research Questions

As we conducted our research we kept the following research questions in mind:

- RQ1: To what extent does StatPlayground afford the collection of data using InfoViz?** Can we eliminate the use of control panels and labels to convey information? Is the user able to check whether the assumptions are satisfied? Is the user able to tell when the difference between means is significant or not? Is the user able to tell the effect size from the design?
- RQ2: To what extent does foreshadowing in StatPlayground motivate exploration?** Is the user able to figure out when the effect size changes categories as the dataset changes?
- RQ3: To what extent are features in StatPlayground discoverable?** Does the user know when a feature can be directly manipulated? Is the user able to figure out how to finely control the properties in a dataset?
- RQ4: To what extent do the expected interactions with StatPlayground match the user's expectations?** Does the user know how to manipulate the data? Does the user know how to lock a property? Does the user know how to set the boundaries on a property? Does the user interact with StatPlayground as expected?



## 1.5 Outline

This thesis is organized as follows:

- **Chapter 2.** We discuss in further detail the research that has investigated the problem with statistics in HCI. We also discuss how this problem has been addressed in the past, and the techniques from other contexts that have inspired our contributions to StatPlayground.
- **Chapter 3.** We describe the iterative process by which we implemented StatPlayground. After drafting paper prototypes and software prototypes, we conducted preliminary studies to receive feedback on the design techniques we used. At the end of the chapter, we discuss how the feedback given from our participants influenced our later designs.
- **Chapter 4.** We describe the final prototype for StatPlayground, with which the user interacts to explore the relationships between sets of data.
- **Chapter 5.** We describe how StatPlayground was evaluated using the think-aloud protocol. We summarize which designs and features were intuitive for our participants and which designs require further work.
- **Chapter 6.** We conclude our findings and make suggestions to further realize StatPlayground's potential.
- **Appendices.** Supplementary information about the implementation, investigation, and user studies.



## Chapter 2

# Related work

In this chapter, we discuss the different approaches made to address the problem of inadequate statistical literacy. We follow up by describing the original version of StatPlayground from Subramanian and Borchers [2017], and describe three research projects that were the main sources of inspiration for our contributions.

### 2.1 Addressing Inadequate Statistical Literacy

Cairns suggested that, to prevent errors in statistics, HCI should follow a standard, similar to how psychology follows the APA Manual for NHST [Cairns, 2007, Association et al., 1994]. However, due to the broad nature of HCI, Thimbleby [2004] suggests that forcing a standard in HCI opposes growth its research. The broad nature of HCI is supported in Cairns' findings: in his study, 41 papers used inferential statistics, but the remaining 39 papers, which were also reviewed, did not.

Applying standards for statistics would oppose growth in HCI.

Some researchers have tried to resolve this problem at a pedagogical level. Utts [2003] found out what concepts in statistical analysis are typically misunderstood and Garfield [1995] proposed a set of principles to improve how

Approaches made to solve problem at a pedagogical level and at a visual level

these concepts are taught. Another way of addressing the problem of inadequate statistical literacy has been to incorporate the use of simulations to explain some abstract concepts [Lane and Peres, 2006]. Simulations can be beneficial, because they make the learning experience more engaging, however, they run the risk of making the student a passive observer.

StatPlayground includes the checking of assumptions, since NHST depends on particular criteria in order for the resulting statistics to bear any value [Cairns, 2007]. For t-test and ANOVA, normality is one assumption, as well is the homogeneity of variance.

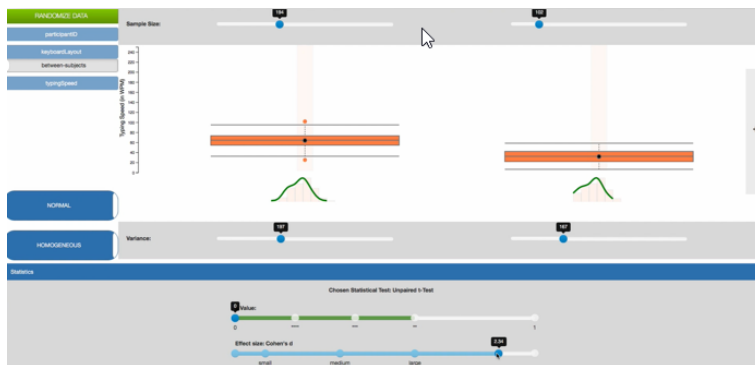
### **2.1.1 StatPlayground: *Subramanian and Borchers, 2017***

Subramanian and Borchers [2017] placed the groundwork for our contributions. They created a version of StatPlayground that allows direct manipulation of the mean, median, and outliers. It presents each distribution as a boxplot with an accompanying histogram. Coloured switches show the user whether assumptions (e.g., normality and homogeneity of variance) are satisfied. StatPlayground follows the principle that students learn by constructing knowledge Garfield [1995]. Giving the user the ability to explore data is also a way to combat errors in statistics [Zuur et al., 2010].

## **2.2 Inspiration for Design Techniques**

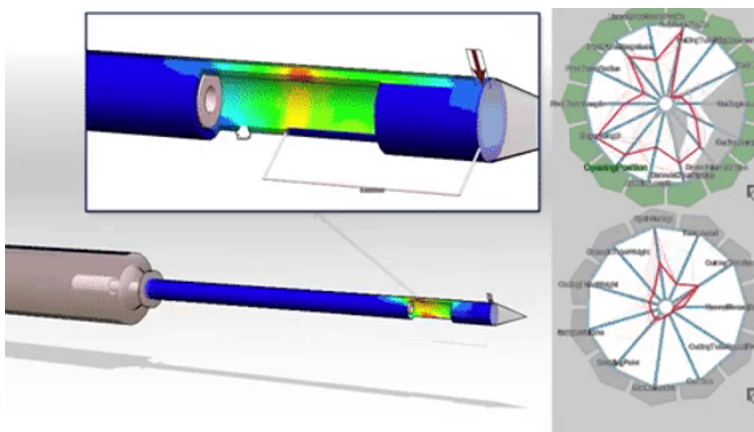
### **2.2.1 Design by Dragging: *Coffey et al., 2013***

Applied in the field of engineering, Design by Dragging is a technique that allows designers to explore the multiple possible configurations to consider when developing a product, such as a medical biopsy device [Coffey et al., 2013]. Design by Dragging allows for both forward and in-



**Figure 2.1:** A screenshot of StatPlayground from Subramanian and Borchers [2017]

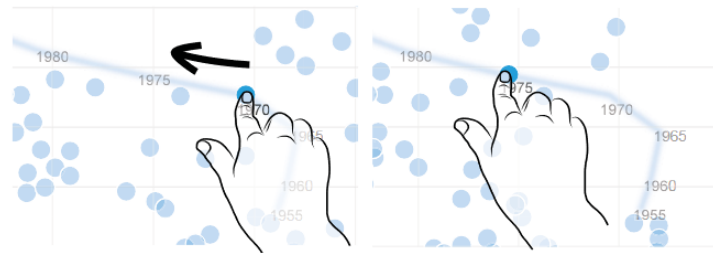
verse design. For forward design, the user clicks and drags on different parts of the simulated product. The location at which the user clicks and the direction of the drag are taken into account when determining which properties are changed. Inverse design uses the same visual space as that in forward design, but instead, as the user clicks and holds down the mouse, visual previews of other possible configurations are displayed close to the cursor. Dragging the cursor to a preview results in changes to the simulated product according to the properties corresponding to the preview. A widget visualizes the different properties that are affected by design changes, and the user may use the widget to lock certain properties at particular values.



**Figure 2.2:** Design by Dragging uses direct manipulation and allows users to lock parameters

### 2.2.2 DimpVis: Kondo and Collins, 2014

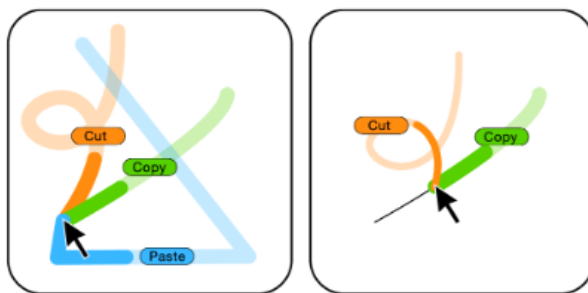
DimpVis provides an interaction technique that applies direct manipulation of visualization for navigating through data throughout time [Kondo and Collins, 2014]. In the example of scatter plots, a user touches a data point and a highlighted path appears to foreshadow its different values in a sequence.



**Figure 2.3:** DimpVis uses foreshadowing to visualize the movement of data points through time. The change in the dataset over time is triggered by the direct manipulation of single data points.

### 2.2.3 OctoPocus: Bau and Mackay, 2008

OctoPocus uses foreshadowing to teach users what gestures that can be performed by clicking and dragging the mouse in predefined sequences [Bau and Mackay, 2008]. When the user holds down the mouse, a series of coloured paths appear, stemming from the position of the cursor. As the user drags the mouse in a particular path, those that become less likely (i.e., the cursor moves in a direction opposite to that of a gesture) fade and eventually disappear, while gestures that become more likely become more visible. Fig. 2.4 illustrates how OctoPocus works for three possible gesture commands.



**Figure 2.4:** Foreshadowing used in OctoPocus gives a preview of possible gestures and fades hints that become less likely as the user moves the cursor





## Chapter 3

# Design Approach

In this chapter, we discuss the process by which we developed StatPlayground. Following the DIA cycle, we started with low-fidelity prototypes ran through iterations of designing, implementing, and analyzing. An online questionnaire was filled out by targeted users, and user studies allowed us to evaluate what designs were preferred and what designs need to be changed.

### 3.1 Iterative Design

We developed our prototype according to the Design-Implement-Analyze (DIA) cycle. The DIA cycle is an iterative process, which starts with a low-fidelity prototype. This prototype is evaluated, and its findings are used to design the next, higher-fidelity prototype. The cycle of design, implement, and analyze continues with increasing fidelity until the final product is complete. In StatPlayground, we started with paper prototypes and then graduated to flip-book prototypes before implementing our software prototypes. Low-fidelity prototypes (paper and flip-book) were evaluated by an expert in statistics and usability design. The software prototypes were evaluated by target users of StatPlayground. The final design is described in Chapter 4.

StatPlayground was developed by following the DIA cycle.

## 3.2 Design Principles

Throughout the design process, we focused on the following principles:

- **Remove the need for control panels.** We want to minimize the amount the user needs to remember about the interface itself, so that more cognitive effort can go into the motivation of StatPlayground. Placing controls directly on the item to be controlled is more effective than using control panels, which are indirect [Norman, 2013, p.155].
- **Only show information when it is needed.** We want to minimize the need for text and instead use symbols, which, when used correctly, convey information faster and are easier to remember.
- **Hide calculations from the user.** The user does not need to know what formulas are used for calculating the p-value nor effect size, since the goal is that the user develops an intuitive understanding of statistics.

## 3.3 Pre-design Decisions

### 3.3.1 Representing Summary Statistics

Boxplots are used to represent the mean, median, variance, and outliers.

The original version of StatPlayground uses boxplots to represent its datasets [Subramanian and Borchers, 2017]. Boxplots are a suitable representation for summary statistics, since they communicate the information that is needed to perform inferential statistics.

In the case of comparing two independent dataset, we determine the p-value using the results of the  $t$ -test. The following equations are used to carry out a two-sample  $t$ -test:

Equations for a two-sample  $t$ -test, which is used for determining the p-value

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (3.1)$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_{\bar{X}_1}^2 + (n_2 - 1)s_{\bar{X}_2}^2}{n_1 + n_2 - 2}} \quad (3.2)$$

Here,  $\bar{X}$  is the mean,  $n$  is the sample size, and  $s^2$  is the variance.

To determine the effect size, we use Cohen's  $d$ , which has the following formula:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{SD_{pooled}} \quad (3.3)$$

where

$$SD_{pooled} = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}} \quad (3.4)$$

Different statistical tests perform different calculations, but the equations above show that to explore the different results in a  $t$ -test, the only properties the user needs to control are the mean, variance, and sample size. All of these properties, except for sample size, can be represented using a boxplot. Matejka and Fitzmaurice [2017] showed that individual data points,  $X_j^i$ , can vary greatly and still produce the same summary statistics, so this information can be hidden for the most part (unless they are outliers).

Equations for Cohen's  $d$ , which is used for determining effect size.

The above calculations reinforce that boxplots communicate the relevant properties in hypothesis testing.

### 3.3.2 Representing Assumption Checks

The original version of StatPlayground uses a coloured histogram to illustrate whether the datasets are normally distributed. We maintained this design, and placed histograms next to their respective boxplots to show that they represented the same data.

Representation of the assumption checks was designed after our preliminary study, so these components were not evaluated until the final evaluations. This mistake is likely responsible for the design's failure during final evaluation, which we will discuss in Chapter 5.

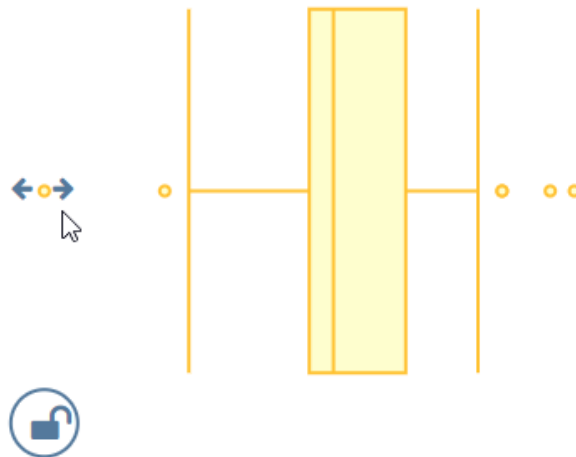
A similar design used for normality testing in the original version of StatPlayground was used.

Assumption checks suffered from insufficient DIA iterations.

### 3.4 Preliminary User Study

We ran the preliminary user study in two separate phases of the DIA cycle. The first phase focused on the implementation of fine-control menus. The second phase focused on choosing a design for foreshadowing, and, unexpectedly, visualising results. An audio recording was made of each session, and users were encouraged to think aloud. The time spent with each user ranged from 15 minutes to 24 minutes.

#### 3.4.1 Fine Control of Properties

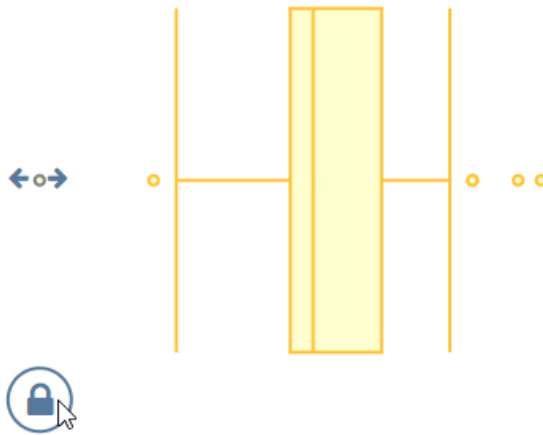


**Figure 3.1:** Fine-control menu prototype. The menu opens then the user hovers over its respective property.

We determined the participants' expectations, then asked for feedback on our design

Figures 3.1, 3.2, and 3.3 illustrate the early prototype of the fine-control menu. The objective of this user study was to determine whether the designs used for each part of the fine-control menu were suitable signifiers for their respective features.

We used the following structure when conducting user studies for evaluating the fine-control feature:

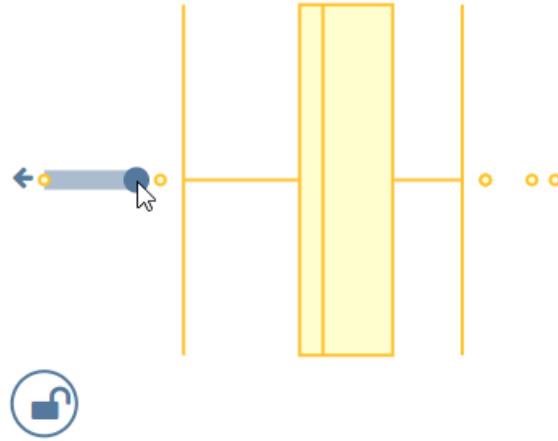


**Figure 3.2:** Fine-control menu prototype. When the property is locked, the lock icon changes the the property turns grey.

1. Present the participant with the design, and allow the user to explore the interface.
2. Ask the participant how they would expect features to be implemented:
  - (a) Fine controls, bounds: "How would you set a minimum or maximum value of an outlier?"
  - (b) Fine controls, locking: "How would you lock an outlier?"
  - (c) Simultaneous manipulation: "How would you modify both the mean and variance at once?"
3. Show the participant what the expected interactions were.
4. Ask the participant for feedback.

All participants interpreted the lock button as expected, however one participant was sure what property was being locked. This confusion is likely the result of placing the lock button too far away from the property used in the user study (the outlier). A suggestion, from another user, to move the lock button closer to the outlier supports our assumption.

The lock feature was generally understood by users.



**Figure 3.3:** Fine-control menu prototype. When an upperbound or lowerbound has been set, a blue circle indicates the maximum/minimum value and the region between the property and the circle indicates the allowed range of the property.

The signifiers for setting upperbounds and lowerbounds needed improvement.

None of the participants interpreted the signifiers for setting upperbounds and lowerbounds as expected. One participant expected that clicking the upperbound button would add data points to the dataset. The rest of the participants expected that the arrows were meant for moving the property itself left and right.

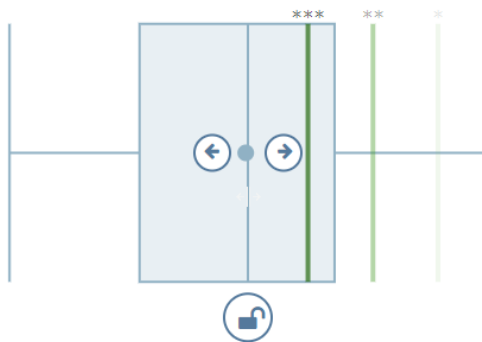
### 3.4.2 Foreshadowing of InfoViz

We came up with four different designs to represent foreshadowing in StatPlayground, and had four participants test each of them. The order in which each participant interacted with the different designs was different, so as to reduce bias from the learning effect.

Foreshadowing design, which uses labels and fading interest points

**Design 1: Vertical Lines** In this design (Fig 3.4), the user clicks and drags on the mean or the median, and a series of vertical lines appear in front of the selected boxplot. Aster-

isks are used as labels to indicate effect size (\* being small, \*\*\* being large). Each line becomes more opaque when the mean or median is closer to the line's value. The lines become more translucent when the mean or median is farther and therefore less likely. The fading behaviour is inspired by the similar technique used in OctoPocus [Bau and Mackay, 2008]. All the lines are the same colour.



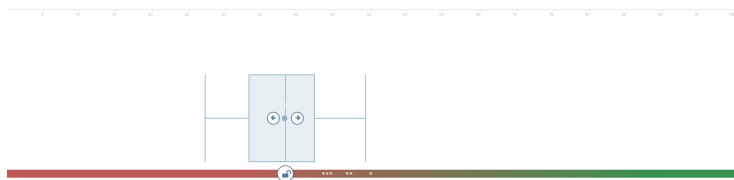
**Figure 3.4:** Prototype for foreshadowing effect size using vertical lines to indicate interest points

Although not the top pic among the participants, this design received positive reviews due to its clarity.

Participants liked that the design was clear

**Design 2: Horizontal Gradient** In this design (Fig 3.5), while the user holds the mouse down, a gradient bar appears below the selected boxplot. Red indicates that the effect size is smaller than "small"; green indicates that the effect size is large. Interest points are marked using asterisks.

Foreshadowing design, which represents Cohen's  $d$  as a gradient



**Figure 3.5:** Prototype for foreshadowing effect size using a gradient

This design had mixed reviews	Reviews were mixed for this design. One participant ranked this design as their favourite among the four, but preferred to see regions to indicate effect size rather than a gradient to represent Cohen's <i>d</i> .
No foreshadowing, but show the current effect size	<b>Design 3: Persistent Box Plot Colours</b> In this design, no foreshadowing is used. However, the user is always aware of the effect size, as the colour of all boxplots change with respect to the effect size. In this prototype, we used arbitrary colours to indicate effect size.
Most participants preferred this design	3 out of 4 participants preferred this design the most. Two of the participants expressed that the arbitrary assignment of colours confused them, implying that meaningful assignment of colours was needed.
No foreshadowing; show the effect size only as the user is moving the data	<b>Design 4: Preview Box Plot Colours</b> In this design, no foreshadowing is used. Moreover, the user does not know what the effect size is, until they click and drag on the boxplot. The selected boxplot changes colour while it is selected, and returns to the default colour when the user lets go of the boxplot. We used arbitrary colours to indicate effect size.
None of the participants preferred this design	All four participants rated this design as their least favourite. Two of the four participants did not understand what the change in colours meant. The other two who understood the behaviour did not like the implementation and preferred to have the boxplot colours to be persistent.
<b>3.4.3 Questionnaire</b>	
An online questionnaire tested different prototypes for experimental design	We wanted to determine whether using colours as an adequate way of conveying experimental design, without the need for using arrows as indicators. Images of the online questionnaire can be found in Section B.1. The questionnaire gave a scenario of a study that could be carried out either using between-subjects design or within-subjects design. We provided four designs:



1. Between-subjects design, using black icons with arrows as indicators
2. Between-subjects design, using coloured icons as indicators
3. Within-subjects design, using black icons with arrows as indicators
4. Within-subjects design, using coloured icons as indicators

First, the participants were asked to interpret the four different designs, to determine whether they were able to tell whether a design was indicating between-subjects design or within-subjects design. Then, the participants were asked, for each type of experimental design, which method of indicating experimental design was preferred (black icons with arrows or coloured icons). The terms "between-subjects" and "within-subjects" were not used in the questionnaire; instead the users interpreted the designs using the example provided. Participants were also asked why they chose their preferred designs.

We tested whether participants were able to interpret the designs, then asked which designs they preferred more.

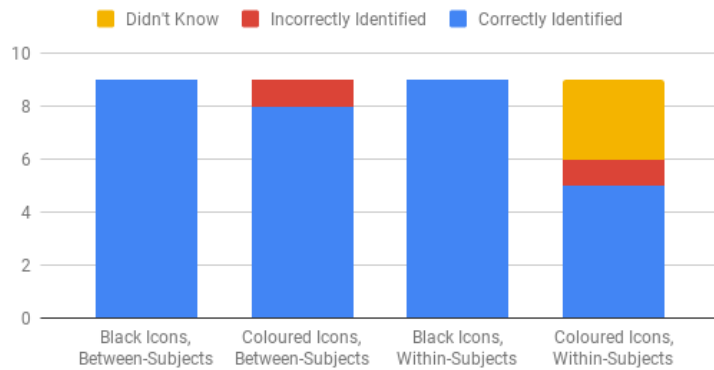
## Results

Figure 3.6 shows how participants interpreted each design. All users correctly interpreted the designs that used the black icons with arrows as indicators. One user incorrectly interpreted the design for between-subjects design using coloured icons. Users had the most trouble interpreting the design for within-subjects design using coloured icons, with only 5 correct interpretations. 1 participant interpreted incorrectly, and 3 participants didn't know how to interpret the design.

Coloured design was largely misinterpreted for within-subjects design.

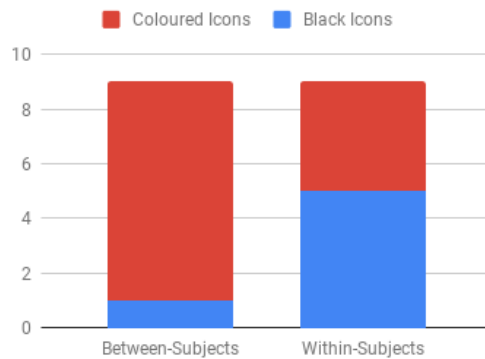
Figure 3.7 shows which designs the participants preferred. For the between-subjects design, participants preferred the design with coloured icons (1 preferred black icons, 8 preferred coloured icons). For within-subjects design, participants were split on their preferences (5 preferred black

Participants preferred the coloured design for between-subjects design



**Figure 3.6:** User interpretation of the different indicators used to differentiate experimental designs.

icons, 4 preferred coloured icons). Quotes on the participants' reasonings for their choices can be found in Section B.1.1.



**Figure 3.7:** User responses to "Which design did you like more?" for both between-subjects design and within-subject design.

We needed to redesign the representation for within-subjects design.

From the results, we determined that our design was suitable for representing between-subjects design using colour, however we needed to update our design for within-subjects using colour. From the feedback provided by the participants, we determined that it was better to use a single colour to represent the population, rather than change the convention of using a different colour per group in between-subjects design to using a different colour per individual in within-subjects design.

### 3.4.4 Participants

6 participants, ages 25-32 (3 females) volunteered to use the interactive prototype. All participants had either some or very little knowledge in statistics. All participants understood terms from summary statistics, however terms such as *p-value* and *effect size* were not universally understood and had to be explained. Only 4 participants evaluated the different designs for foreshadowing.

6 participants evaluated the low-level prototypes

The questionnaire was filled out online by 9 participants within the age range of 20-35. The exact ages and genders were not recorded. All participants had either some or very little knowledge in statistics.

9 participants provided feedback on experimental design prototypes

## 3.5 Abandoned Feature: Simultaneous Manipulation

In the beginning of our research, we hoped we could add the ability to manipulate multiple parameters (e.g., mean and variance) simultaneously. There are tools that enable simultaneous manipulation of multiple parameters via touch screens [Nielsen et al., 2016, Coffey et al., 2013], however, the scope of StatPlayground is limited to WIMP interactions. While drafting paper prototypes in the early iterations of the DIA cycle, we struggled to design a solution that had a natural mapping of gestures to property manipulation. In our preliminary user study, asked participants, "How would you modify both the mean and variance at once?" to see if they could come up with a solution that would make sense to them. One suggested the use of keyboard shortcuts, as implemented in Adobe Illustrator<sup>1</sup>, but followed up by saying that it is sufficient to modify the mean and variance independently. The remaining participants were not able to come up with a solution. Since users from the preliminary user study did not show a direct interest or need for simultaneous manipulation, this feature

We decided not to implement simultaneous manipulation after a lack of interest shown by user.

<sup>1</sup><https://helpx.adobe.com/illustrator/using/default-keyboard-shortcuts.html>

was abandoned.

### 3.5.1 Resulting Changes

We made the fine-control buttons more uniform and made sure they were close to their subjects.

Feedback from participants was carefully taken into consideration. For the fine-control menu, we fit the upperbound and lowerbound icons into a button, similar to that of the lock button. By using the same button style, the expectation is that the user would conceptually cluster the upperbound/lowerbound buttons and the lock buttons together. We also moved the lock button closer to the outliers and mean, which are represented as circles. The position of the lock button stayed for the taller boxplot properties, such as the median and whiskers.

We combined the participants' most preferred designs.

We implemented a combination of Design 1 and Design 3 for adding foreshadowing. The boxplots will always have a colour that indicates the current effect size, and a set of vertical bars appear during click-and-drag to signify effect size interest points. We assigned meaningful colours to represent the varying levels of effect size.

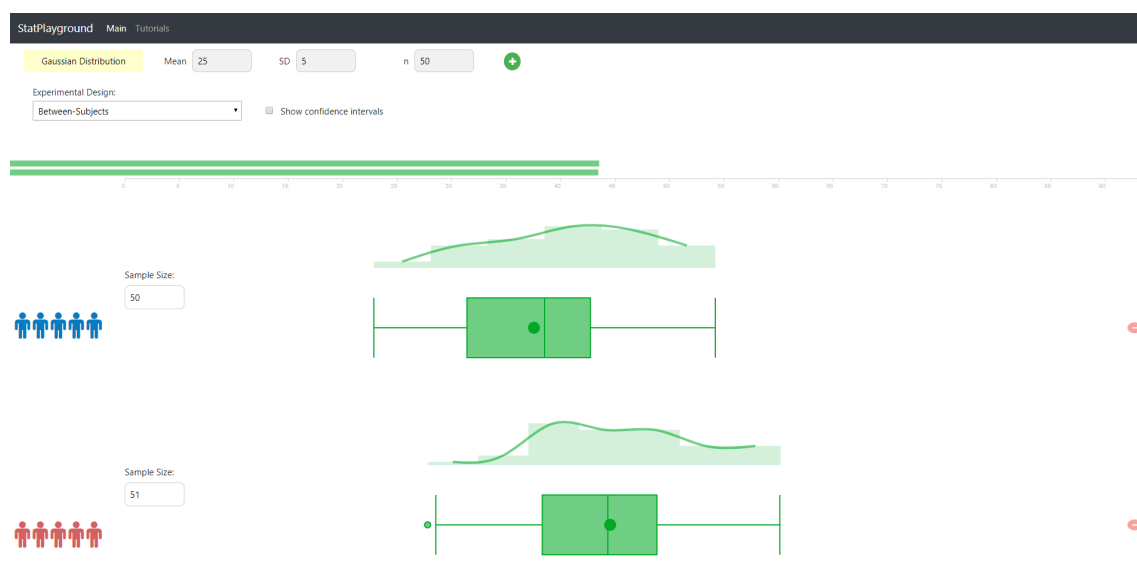
We made adjusted the prototype used for representing within-subjects design, however, this feature was not further evaluated in this thesis, and requires more attention in the future.

## Chapter 4

# Interaction Design

In the previous chapter, we discussed how we developed StatPlayground according to the DIA cycle. In this chapter, we describe the final interactive design of StatPlayground.

### 4.1 Design Layout of StatPlayground



**Figure 4.1:** Screenshot of StatPlayground with two datasets

The layout of StatPlayground is divided into three main sections: the dataset configurations, the series of distributions, and the statistical results.

In the dataset configurations, the user can add distributions to the dataset, determine the experimental design, and choose whether the confidence interval is visible or not.

The area for the series of distributions is the main focus of the StatPlayground. It is where the user is able to manipulate the dataset and use the design strategies to infer the results of the statistical analysis.

The bottom section shows the statistical results.

## 4.2 Configuring the Dataset

How to create a dataset

To create a dataset, the user adds distributions to the working space by clicking the '+' icon to add the distribution as a boxplot to the dataset. If the user wishes to configure the distribution before adding it to the dataset, they may select the type of distribution from a drop-down menu. The user may also set the mean, median, and sample size. These configurations will be used to influence where data points will be created within the distribution.

How to show/hide the confidence interval

The user has the option of viewing the confidence intervals of each distribution checking or un-checking the respective field. By default, the confidence interval is hidden. The confidence interval is represented as a line segment that hovers in front of the boxplot.

How to set the experimental design

A drop-down list is used to select whether the experimental design is between-subjects or within-subjects. The selection of the experimental design is symbolically represented via groups of person-shaped icons to the left of each distribution. For between-subjects design, each set of icons is assigned a different colour. This represents the concept that different groups of people are associated with each distribution. For within-subjects design, all sets of icons are the

same colour. This represents the idea that all subjects have taken part in all treatments.

## 4.3 Manipulating Properties

### 4.3.1 Changing Values

The boxplots can be directly manipulated by clicking and dragging on particular parts on the horizontal axis. Properties that can be directly manipulated are the mean, median, and outliers. The variance can also be manipulated by clicking and dragging on the end of either whisker. When the user hovers the mouse cursor over these properties, a tooltip appears to show the name of the property and its corresponding value. Additionally, the cursor changes in style from Cursor A in Fig 4.6 to Cursor B to indicate horizontal movement. to indicate to the user that they can perform a click-and-drag action. Fig 4.2 illustrates a boxplot might look when the user hovers over the median.

How to manipulate the mean, median, variance, and outliers

The user can turn a data point into an outlier with a single click anywhere outside the boxplot on the centre axis. Outliers may also be removed by clicking them and dragging them within the boxplot region.

How to create/delete an outlier

The sample size is presented as an input field to the left of each boxplot. Its value can be changed by either typing a numerical value, or by using the arrow buttons inside the input field.

How to edit the sample size

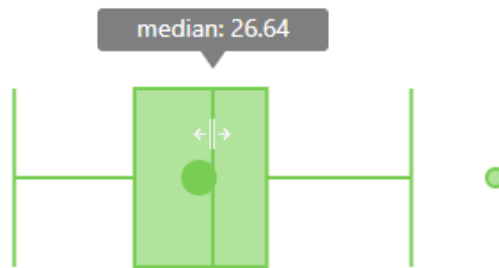
### 4.3.2 Fine Control of Properties

The user can constrain properties by opening the fine-control menu (Fig 4.3). The fine-control menu opens when the user double-clicks on the mean, median, whiskers, and outliers. The menu is presented as three circular buttons.

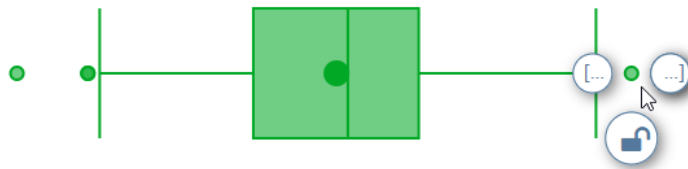
How to trigger the fine-control menu

One button has a lock icon and, when clicked, prevents the

How to lock a property



**Figure 4.2:** A tooltip appears when the cursor hovers over a property (e.g., the median), indicating what property is being hovered and its current value. The cursor style changes to indicate that the property can be clicked-and-dragged sideways.



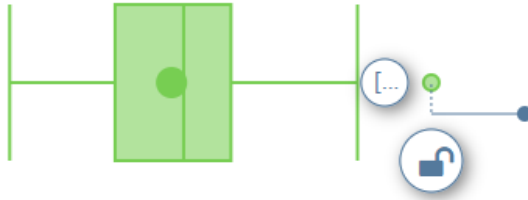
**Figure 4.3:** The default fine-controls menu, shown when the user double-clicks on a controllable property.

property from moving. When the property is locked, it turns grey to indicate to the user that it cannot be moved until it is unlocked. Cursor C in Fig. 4.6 is used when hovering over the lock button to indicate that the button is meant to be clicked.

How to set a value  
upperbound or  
lowerbound

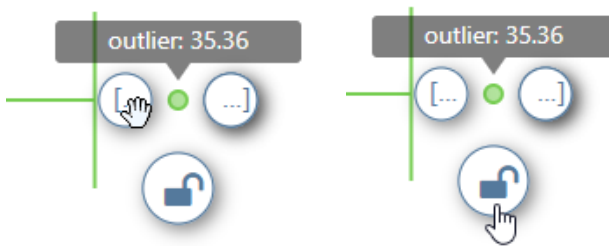
The other two buttons are reserved for setting the upperbound and lowerbound of the selected property. The text within these buttons follow the mathematical notation of '['...' and '...]' to indicate the lowerbound and upperbound respectively. When the user clicks and drags one of these buttons, the button transforms into a small circle with a line that connects back to the property. The circle represents the minimum or maximum value, and the line represents the available range of the property. Fig. 3.3 shows what the fine-control menu looks like when a boundary has been set. To unset a boundary, the user clicks and drags the circle back to the current value of the property. Cursor D in Fig.





**Figure 4.4:** The fine-controls menu, where the upperbound has been set. The user will not be able to move the set property beyond the blue circle.

4.6 is used to indicate that the user can drag the button. Fig. 4.5 shows the different cursor styles applied.



**Figure 4.5:** Cursor styles used for both boundary buttons (left) and lock button (right).

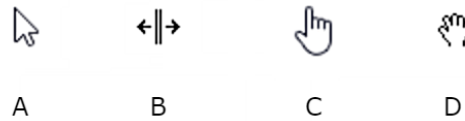
## 4.4 Performing Assumption Checks

A histogram of the distribution is shown above each boxplot (Fig. 4.7). Each histogram will either be red or green, depending on whether the histogram is normally distributed. A red histogram indicates that the data is not normally distributed. A green histogram indicates that the data is normally distributed.

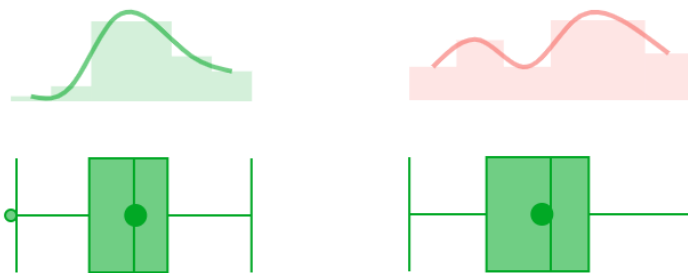
How to check for normality

When the a new distribution is added to the dataset, a horizontal bar appears in front of the boxplot, and is animated to move to the top-left corner of the dataset space (Fig. 4.8). This horizontal bar represents the variance of the distribution. These horizontal bars allow for the user to get an

How to check for homogeneity of variances



**Figure 4.6:** The different cursors styles used in StatPlayground. Cursor A is the default style. Cursor B indicates that a boxplot property (e.g., the mean) can be clicked and dragged horizontally. Cursor C indicates that a button (i.e., the lock button) can be pressed. Cursor D indicates that a button (e.g., the lower bound button) can be clicked and dragged.



**Figure 4.7:** Left: The histogram is green when the dataset is normally distributed; Right: The histogram is red when the dataset is not normally distributed.

overview of the variances of all boxplots. The horizontal bars will all be green or red, depending on whether the assumption of homogeneous variances is met or not met, respectively.

Tooltips act as discoverable labels

When the user hovers the cursor over a histogram, a tooltip appears to indicate what assumption is being checked, and whether the assumption is met or not, (e.g., "Normality Test: Pass"). Likewise, when the user hovers the cursor over the horizontal bars to represent the variances, a tooltip appears to indicate whether the variances are homogeneous (e.g., "Homogeneity of Variance Test: Fail").



**Figure 4.8:** Left: Three bars representing the variances of three datasets are green when the assumption of homogeneity of variances is satisfied; Right: The bars are red when the assumption is not satisfied.

## 4.5 Viewing the Results

The results in text-form are shown at the bottom of the page and are updated in realtime. The results section tells you the effect size and p-value. StatPlayground uses Cohen's d to determine the effect sizes: small, medium, and large. The assignment of effect size to Cohen's d is shown in Table 4.1

The effect size and p-value are printed in simple text form

The exact p-value of the statistical results is not presented to the user. Instead the user views the p-value as either " $< 0.05$ " or " $\geq 0.05$ ".

Effect Size	d
Small	0.20
Medium	0.50
Large	0.80

**Table 4.1:** Effect sizes and their corresponding Cohen's d value

The results of the statistical analysis are reflected in the colour of the boxplots in the dataset (Fig. 4.9). Regardless of the effect size, if the p-value is greater than or equal to the threshold, 0.05, all boxplots will be red. If the p-value is less than 0.05, then the boxplots will be assigned the colours yellow, yellow-green, and green for effect sizes: small, medium, and large, respectively. Table 4.2 summarizes the colours associated to the different combinations of results.

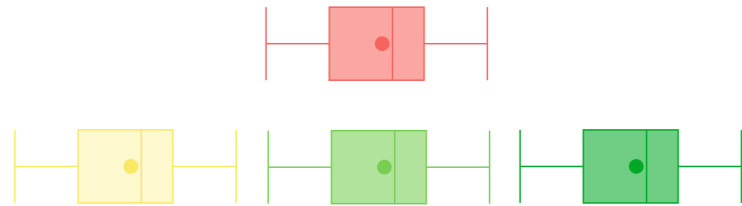
Different colours represent different p-values and effect sizes

The colours for the varying effect sizes are also used as foreshadowing for the user (Fig. 4.10). When the user holds

The same colours are used to foreshadow interest points for effect size

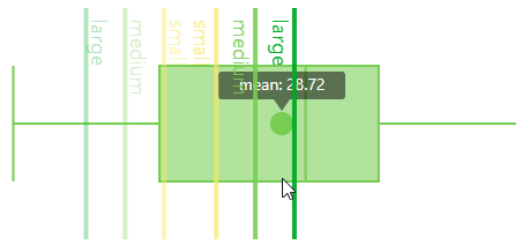
Effect Size	p-value	Colour in StatPlayground
Small/Medium/Large	$\geq 0.05$	Red
Small	$< 0.05$	Yellow
Medium	$< 0.05$	Yellow-Green
Large	$< 0.05$	Green

**Table 4.2:** Table of colours used in StatPlayground assigned to each combination of effect size and p-value



**Figure 4.9:** Top: The colour of a boxplot when  $p \geq 0.05$ ; Bottom, left-to-right: Yellow boxplot for small effect size, yellow-green boxplot for medium effect size, and green boxplot for large effect size.

down the mouse, as they click and drag the mean or median, a series of vertical lines appear in front of the selected boxplot. These lines represent the interest points at which the effect size and p-value would change. The lines are assigned the same colours as the colours assigned to the different effect sizes: yellow, yellow-green, and green.



**Figure 4.10:** Foreshadowing of interest points are projected on top of the boxplot. Interest points that are closer to the selected property are more opaque than those that are farther away, and therefore less likely.

## Chapter 5

# Evaluation

In this chapter, we describe how the later versions of Stat-Playground were evaluated. The evaluation was run in two phases. In the first phase, we collected data from 8 participants, however the audio data for two participants failed to record. As a result, some changes were made with according to the feedback from the remaining 6 participants. We also adjusted the format at which the evaluation was held. We then ran another phase of user studies with seven participants.

### 5.1 User Study: Phase 1

#### 5.1.1 Format

We asked the participants how much they knew or remembered about statistics. Slides from the lecture that all students had attended were available for the participants' reference. We verbally explained different terms, such as *variance*, *effect size*, and *p-value*. After explaining the terminology, we presented the user the interface, which had already been set up with two boxplots. We described the same scenario to each participant, and pointed out on the interface that each boxplot represented each group. Before providing the user with a list of tasks, we explained some of the

basic components. We explained that the user can click and drag on different parts, and excluded the new features (i.e., foreshadowing, use of colour, fine controls) from our explanation. We used the think aloud protocol in the study, and encouraged the participants to verbalize their thoughts as they were asked to complete each task. After the tasks were performed, the evaluator asked the user for their general feedback feedback.

### 5.1.2 Tasks

We observed whether the user was able to perform the following actions:

1. Move the data by clicking on mean
2. Move the boxplot to a position where the null hypothesis cannot be rejected
3. Move the boxplot to a position where the null hypothesis can be rejected
4. Determine whether the assumption check for homogeneity of variance is satisfied
5. Determine whether the assumption check for normality is satisfied
6. Lock an outlier
7. Set an upperbound on an outlier

### 5.1.3 Participants

8 participants (two females) took part in the study. All participants are university students and had at some point attended a lecture that covered introductory statistics. To minimize the learning effect, participants were not allowed to have participated in the preliminary study. Due to technical failure, the data from two participants could not be used.

### 5.1.4 Limitations After Phase 1

We decided our research required another iteration of the DIA cycle for three main reasons.

Firstly, it was discovered that the audio was not captured for two participants meant that only data from six participants was usable for analysis. Although the screen was able to capture the participants' mouse movements, without audio, it was impossible to interpret when the participants tried to click or what motivated their mouse movements.

We were only able to collect data from 6 out of the 8 participants.

Secondly, the version of StatPlayground used for this evaluation still contained many bugs and missing implementation. For example, users would be able to create an outlier, but the boxplot would freeze after doing so. Participants had to refresh the application each time this would happen, and the need to refresh the application was a distraction for the participants. In another example, the histograms would sometimes be green, indicating that the data is normally distributed, when it was clear that the distribution was heavily skewed or bi-modal. A similar problem appeared for the test for homogeneity of variance. These bugs made it difficult to make sense of the colours used. In an attempt to work around this problem, we asked the participants to imagine that the correct colour was assigned.

Bugs interrupted the cognitive process and made it difficult to evaluate some features.

Thirdly, the evaluation itself had several problems. Information was only communicated to the participants verbally, or by pointing at the screen. This required that participants used their memory to make sense of the prototype and the example scenario provided. Forcing participants to remember information created a distraction when asking them to perform the different tasks. Moreover, there was a lack in structure with how the investigator communicated with the participants. The investigator was inconsistent with how instructions were delivered; some participants were asked to figure out how to interact with a feature, whereas other participants were told. A lack of script led the investigator of the user study to overexplain details when participants expressed confusion.

There were problems with how we conducted the user study.

## Resulting changes

Another iteration of the DIA cycle was needed	It might have been acceptable to use the data from only six participants. However, due to the bugs in the implementation that hindered the evaluation process, and due to the lack of structure and consistency in how the evaluation was carried out, it was decided that another iteration of the DIA cycle was needed. A detailed list of changes made after Phase 1 can be found in Appendix B.2.2.
Participants were given "cheat sheets" to eliminate the need to memorize contextual information.	We made sure that information was presented to the participants both visually as verbally. The evaluator presented the participants with a "cheat sheet", which quickly explained statistical terms. It also contained a picture of a boxplot, drawn in the same manner as displayed in StatPlayground, with labels to identify which parts represented the mean, median, variance, and outlier. This was to make sure that all participants were on the same level of knowledge going into the tasks. The scenario description was also printed out, as well as the list of tasks for the participant to perform. During evaluation, the evaluator made sure to read out the tasks word-for-word to ensure consistency with how the questions were being asked. After the tasks were performed, the evaluator asked for feedback in further detail.
Addition of a Tutorial Page	A tutorial was made to replace the need for an instructor. It goes through all the features offered by StatPlayground. This tutorial is created as a separate page on the StatPlayground application, and is the first page the user sees upon opening the application. The tutorial is not interactive, but the user can easily switch back and forth between the tutorial page and the interactive page.
Changes were made to the design based on feedback.	The upperbound and lowerbound buttons were still mistaken for indicators. As suggested by one participant, the arrow icons were replaced with mathematical notations "[..." and "...]", Bugs were also fixed which made features like the assumption checks more testable.



## 5.2 User Study: Phase 2

### 5.2.1 Format

Phase 2 of the evaluation followed the following structure:

1. **Establish a base knowledge.** We prepared a hand-out with the relevant terminology used in StatPlayground and a labelled image of a boxplot (Fig. B.5).
2. **Provide context.** We prepared a second handout, which described the example scenario used for the user study (Fig. B.6). Our example described the comparison of stress induced by two different designs for a web portal. All participants had experience with the web portal used in the example.
3. **Go through the tutorial.** Participants were told to take their time to go through the tutorial and inform the investigator when they were ready. The investigator performed small tasks (e.g., re-organize papers, check mobile phone, grab a drink) to minimize pressure on the participant to finish the tutorial quickly.
4. **Set up the working space.** The investigator asked the participant to create two boxplots, each boxplot represented an alternative design for the web portal. The investigator reiterated the meaning of each boxplot as they were created (e.g., "This boxplot represents the old design").
5. **Perform the tasks to manipulate the data.** The following section lists the different tasks the participants were asked to complete. The tasks were read out verbatim and were presented to the participant as a handout (Fig. B.7).
6. **Collect feedback.** If the participant tried to complete certain tasks that were different from expected, they were asked what their thought process was. Participants were also asked what they found confusing about the design, what they didn't like, and whether they found StatPlayground useful.

### 5.2.2 Tasks

The tasks given to the participant were as follows:

1. How might the dataset look if the new design did not affect the change in heart rate?
2. How might the dataset look if the new design reduces the amount of stress (i.e., minimizes the increase in heart rate) with a medium effect size?
3. Change the distributions so that the following assumptions are both met:
  - (a) Homogeneity of Variance: The spread of all distributions are similar
  - (b) Normality: Each set of data is normally distributed
4. Create an outlier to represent a user that drank espresso before using the new design. This user would have a higher heart rate than the rest of the group.
5. How would you manipulate this outlier, so that they would have the same heart rate regardless of how stressful the user interface is?
6. How would you manipulate this outlier, so that the heart rate will increase by a minimum of 55 bpm, regardless of how stressful the user interface is?

### 5.2.3 Participants

7 participants volunteered, who have never used StatPlayground previously

We collected data from 7 participants (two females), all of which are university students, and have basic experience with statistics. To minimize the learning effect, participants for this user study were not allowed to have participated in the preliminary study nor the first round of user studies.

## 5.2.4 Results

### Quantitative Results

Table 5.1 shows the accuracy at which tasks were performed according to the expected behaviour.

Expected Behaviour	Correct	Incorrect	Accuracy
Moved the data by clicking and dragging the mean or median	5	2	71%
Used the boxplot colour and hints to determine when the null hypothesis could not be rejected	5	2	71%
Used the boxplot and hints to determine where the null hypothesis would be rejected, with a medium effect size	3	1	75% <sup>1</sup>
Used the component to determine whether homogeneity of variance was met	2	5	29%
Used the component to check that normality was met	7	0	100%
Changed the variance using the whiskers	7	0	100%
Created an outlier by clicking in the correct location	5	2	71%
Opened fine controls by double-clicking on the property	3	4	43%
Locked the outlier by clicking the button	6	1	86%
Click-and-dragged the bounds button to set minimum value	4	3	57%

**Table 5.1:** Table of how accurately participants used the application according to the expected behaviour.

<sup>1</sup>Three participants were asked a different question, and were thus excluded from the result

**Moved the data by clicking and dragging the mean or median** The two participants who did not click and drag on the mean or median clicked on the body of the boxplot

**Used the boxplot colour and hints to determine when the null hypothesis could not be rejected** One participant relied on the Results section only One participant did not discover the desired region, and stopped prematurely in the “medium” range

**Used the boxplot and hints to determine where the null hypothesis would be rejected, with a medium effect size** The participant who failed to complete this task relied on the Results section only

**Used the component to determine whether homogeneity of variance was met** The five participants who failed to complete this task did not know where to find this feature.

**Used the component to check that normality was met** All participants correctly completed this task

**Changed the variance using the whiskers** All participants correctly completed this task

**Created an outlier by clicking in the correct location** One participant tried to right-click on the working space to create an outlier One participant did not know how to add an outlier, and when told how to create an outlier, tried to double-click instead of single-click.

**Opened fine controls by double-clicking on the property** Four participants tried to right-click on the property.

**Locked the outlier by clicking the button** One participant had to be told to click the “lock” button

**Click-and-dragged the bounds button to set minimum value** Three participants only single-clicked the button, and did not drag it

### Qualitative Results

In the feedback portion of the user study, the investigator went over any unexpected behaviours produced by the participant, and asked what the user thought of the interface.

**Question: What didn’t you like about the interface, or what was confusing?**

- One user expressed that they didn’t understand what the hints represented. And why the hints were represented as a 2-tail test rather than a 1-tail test
- 2 users would have preferred if the Results section was presented in a location that could always be visible, and didn’t have to scroll down
- One user would like to see vertical lines to show where on the x-axis the value is
- One user would have liked to be able to see precisely what value the upper/lower bounds are being set at
- One user expressed that because all features would use the same shade of red to indicate “incorrectness” and the same shade of green to indicate “correctness”, the features themselves ended up being clustered together
- One user pointed out that the layout for within-subjects design didn’t make sense, in both the interactive application as well as the graphic used in the tutorial

- "I think the colours just sort of confused me a bit, because I think the tutorial doesn't tell you what the yellow means"
- From the user who used the Results section and didn't use the hints or boxplot color to determine the interest points: "At the beginning I was a bit confused with the colours. I thought it was about normality."
- "The colours are helpful, but they are confusing a bit."
- Regarding the hints: "I didn't really understand them. You told me medium, but I don't know what "medium" means."

#### Question: What do you like about the application?

- "I like the simplicity of the tool actually. Usually the other stats tool that you use there you know bunch of buttons and so many things that you tend to get confused. So that's not there, which is good."
- "I like the fact that it looks like an interface an not a table. Which is what most websites use."

**Other Unexpected Behaviour** One user tried to interact with the histograms by clicking and dragging

### 5.3 Limitations

Our evaluations were limited by time and motivation of the participants

Although we implemented features such as confidence interval. Due to time constrictions and limited backend implementation, we did not give the participants enough time to fully explore the interface. Participants who volunteered for our study were not all equally motivated, this is reflected in the amount of time spent in the tutorial (shortest: 1min 45 seconds; longest: 14 minutes). The amount of time

spent on the tutorials was reflected in how much the user remembered.

The positive responses of the participants about the overall design need to be taken with a grain of salt due to the possibility of response bias [Furnham, 1986].

## 5.4 Discussion

In this section, we discuss why participants might have interpreted our design differently than expected using Gestalt laws [Soegaard, 2015, Wulf, 1922].

### 5.4.1 Gestalt Laws

**Gestalt Law of Similarity** In our implementation of Stat-Playground use the colours red to indicate incorrectness or that something was wrong. In the design process, we used red to indicate *failure* to reject the null hypothesis. However, in statistics, failure to reject the null hypothesis should not be perceived as "failure" as it does for the assumption checks.

Red and green did not have the same meaning for different components

**Gestalt Law of Proximity** We placed histograms directly above their respective boxplots because they are associated with the same data. However, when we were discussing only effect size with the participants, they got confused by the colour changes in the histogram. This might have only been problem with the software, since the histogram would recalculate everytime the data points in the distribution moved around. This was also affected by the fact that the test for normal distributions was only pseudo-implemented. And depending on the change in histogram, the result of the test for normality would change between pass and fail when the true shape of the data has not changed (e.g., when moving the mean). Although the data between the histograms and the boxplots are the same, for the sake of exploring data, it would be better to group areas

Changes observed in one object (e.g., location of boxplot) was assumed to also affect nearby objects (e.g., normality check)

in terms of purpose. In addition, because a pseudo function was implemented for the normality assumption check, some changes to the the dataset caused the normality test to pass or fail when nothing should have changed. This sometimes led participants to draw incorrect conclusions. For example, a user thought that, because the normality test suddenly changed while moving the mean, the user assumed that this was the sign for the p-value changing from not significant to significant.

Past experience might have led to different interpretations of colours.

**Gestalt Law of Past Experience** Because in western culture, red tends to be associate with "bad" and green tends to be associated with "good". Something that would be interesting to determine is whether people associate "goodness" or even colour with "similarity". This is something that could be determined using an implicit association test [Greenwald et al., 1998].

Users preferred right-click

The users who right-clicked to open the fine control menus did so most likely because of experience with context menus, whereby the user right-clicks on the mouse for more options.

#### 5.4.2 Shortcomings of Exploration

Shortcomings of think-aloud protocol

Some participants were not the most comfortable with verbalising their thoughts, and required a lot of probing from the researcher. The researcher also noticed that some users got nervous when the users did not understand something. Although we tried to reassure the participants that their knowledge in statistics was not being tested, the task made it difficult for participants to forget this. The benefits of discoverability are reaped when the user has motivated enough to explore the interface. However, not all users might have the patience nor motivation to look for all possible features.

Shortcomings of forcing discoverability



## Chapter 6

# Summary and Future Work

In this chapter, we discuss the findings of our research, its limitations, and proposals for future research.

### 6.1 Summary and Contributions

This topic was inspired by the findings that knowledge in statistics is lacking in the field of HCI, and requires attention to protect the validity of future research. Research in the past has sought to address this problem in several ways. In this thesis, we focused on providing a space for users to play around with data to explore the relationships and make connections. We extended previous research by adding fine control of properties and incorporating information to the workspace using colours and presenting zones to foreshadow statistical results.

We worked in an iterative process, abiding by the DIA cycle. After looking at techniques applied in previous research, we came up with different prototypes to be evaluated by users in a think aloud protocol. From our findings, we tweaked our design and continued to implement our design.

We added fine control of properties and foreshadowing of statistical results to StatPlayground

We worked in an iterative process and received feedback on design options using the think aloud protocol

---

<p>We ran a first set of user studies, but discovered that the data for two users had not been recorded correctly.</p>	<p>To evaluate our design, we ran a user study with eight participants. Unfortunately, technical failure with the laptop used to carry out the study resulted with incomplete recordings from two participants. To validate our design, we needed feedback from more participants, so we decided that another set of user studies was needed.</p>
<p>We made changes to the design, fixed bugs, and ran another set of user studies</p>	<p>Before running the new set of user studies, we made some adjustments according to feedback from participants, and fixed some bugs which made the user study process distracting. In addition, we created a tutorial, so that the user could see how each feature is meant to work before completing the tasks. As well, we made adjustments to the context and questions asked. Changes to the format of the study were also made. The example scenario provided to the participants was changed to one that was well understood and relevant. Participants were also provided with a concise cheat sheet for statistical terms used in the user study, as well as the list of questions. We ran the user study with seven participants.</p>
<p>Foreshadowing and use of colours showed to be helpful, however work still needs to be done for fine controls</p>	<p>Users spent varying amounts of time on the tutorial, and this may have impacted their ability to remember what features and interactions were available. Most participants were able to use the foreshadowing feature and the colours of the boxplots to determine whether the results had a significant p-value and what the effect size was. Users were also able to easily check whether the assumption of normality was satisfied. However, most users did not know how to use the interface to check the assumption of homogeneous variances. Participants also had a tendency to use right-clicking as a mechanism to open the fine-controls menu instead of the expected double-click. While the locking of properties was straightforward for most users, three of the seven users assumed that the signifier for setting upper- and lowerbounds was a single-click button and not a click-and-drag feature.</p>
<p>Participants saw potential in StatPlayground</p>	<p>Users however reported that, overall, they enjoyed the experience of using the tool. They liked that the design was simple compared to other statistical tools they have used in the past.</p>

## 6.2 Future Work

### 6.2.1 Further Testing

The current design of StatPlayground is meant to support ANOVA testing, but this feature still needs to be evaluated by participants.

ANOVA needs to be tested

In the previous chapter, we discussed some of the problems encountered by users with the interface. More iterations of the DIA cycles are needed to further improve the design, so that it is more intuitive for the user. It would be interesting to introduce techniques from social psychology, such as the implicit association test ([Greenwald et al., 1998]), to determine what choice of colours affect how users perceive information.

Our evaluation laid groundwork for further iterations of the DIA cycle

### 6.2.2 Inverse Design

In StatPlayground, a user might want to see what different configurations are possible with a given set of results. For example, if the  $p \geq 0.05$ , there are several ways the data might look; the boxplot means might be closer together with larger variances, or the means might be farther apart with smaller variances. Design by Dragging gave this technique the term "inverse design" [Coffey et al., 2013]. This technique would work well with our fine-control feature, as it would allow the user to isolate the properties they would like to study. Applying the same techniques as used in OctoPocus Bau and Mackay [2008], since not all configurations are equally likely in statistics. For example, it is much more likely to observe similar means with smaller variances than dissimilar means and extremely large variances.

Inverse design is a feature that was not implemented, but might be useful for users



## Appendix A

# Implementation

The prototype used for this thesis was developed using Angular<sup>1</sup> v5.2.7. The evaluations were carried out on a Windows laptop with a 14-inch (16:9) LED display. Video and audio were captured using OBS Studio<sup>2</sup> v21.1.2.

Future updates on the StatPlayground project will be found on the Media Computing Group website<sup>3</sup>.

---

<sup>1</sup><https://angular.io>

<sup>2</sup><https://obsproject.com>

<sup>3</sup><http://hci.rwth-aachen.de/statplayground>



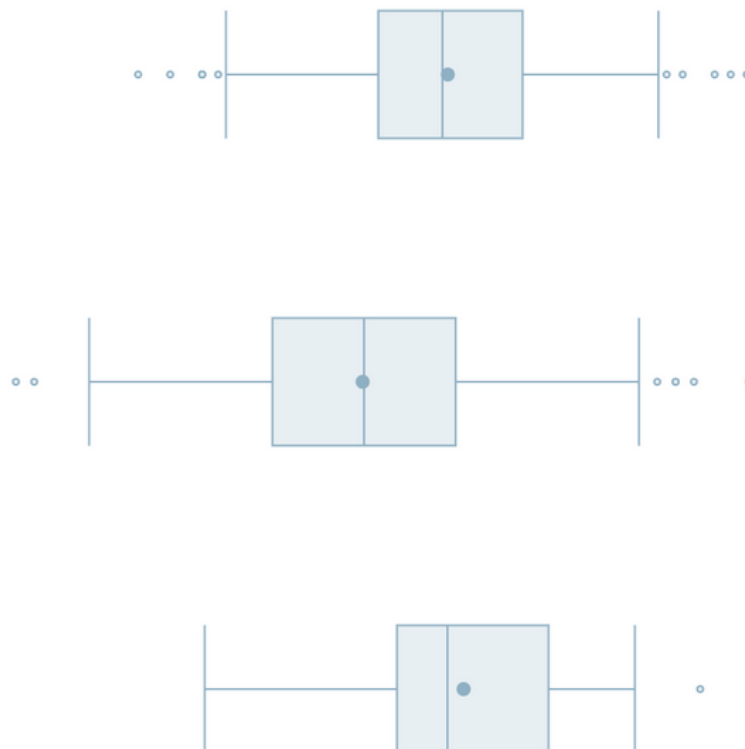
## Appendix B

# User Study

### B.1 Questionnaire for Experimental Design

Figures B.1, B.2, B.3, and B.4 show screenshots of the online questionnaire for testing the different designs for experimental design. The questionnaire was generated using Google Forms.

Lets say we had the scores of many students in three different exams: mathematics, biology, and history. We let each boxplot represent the results for each exam.



### The Situation

The way we collect the test data can be done in two ways:

A) We randomly assign the students into 3 groups. Group 1 takes the mathematics test; Group 2 takes the biology test; Group 3 take the history test.

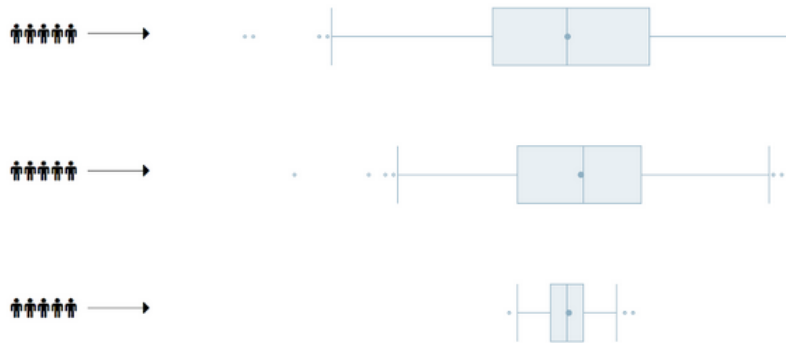
B) All students take all 3 exams

**Figure B.1:** Scenario used for explaining experimental design



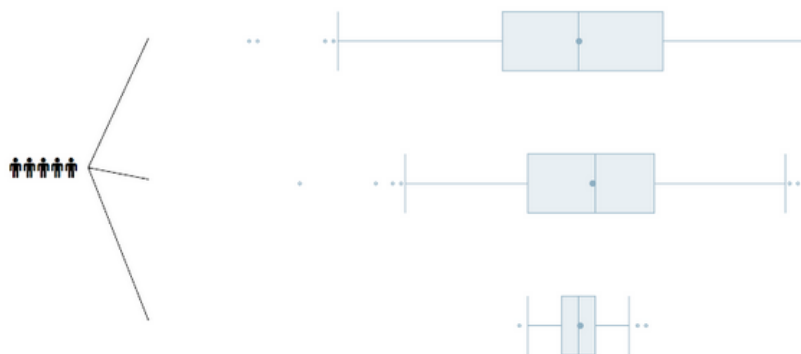
Design 1: Black with Arrows

Based on the design to the left of the box plots (the people icons), which method was used to collect the data? \*



- A) There are 3 separate groups. Group 1 takes the mathematics test; Group 2 takes the biology test; Group 3 take the history test.
- B) All students take all 3 exams
- C) I can't tell from the design

(Same question as above) \*



- A) There are 3 separate groups. Group 1 takes the mathematics test; Group 2 takes the biology test; Group 3 take the history test.
- B) All students take all 3 exams
- C) I can't tell from the design

Figure B.2: Questions to determine how participants interpret lines and arrows to indicate experimental design.

### Design 2: Coloured Icons

Based on the design to the left of the box plots (the people icons), which method was used to collect the data? \*



- A) There are 3 separate groups. Group 1 takes the mathematics test; Group 2 takes the biology test; Group 3 take the history test.
- B) All students take all 3 exams
- C) I can't tell from the design

Based on the design to the left of the box plots (the people icons), which method was used to collect the data? \*

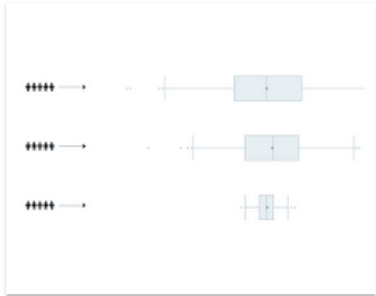
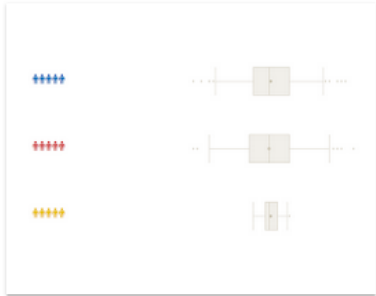


- A) There are 3 separate groups. Group 1 takes the mathematics test; Group 2 takes the biology test; Group 3 take the history test.
- B) All students take all 3 exams
- C) I can't tell from the design

**Figure B.3:** Questions determining how participants interpret colour to indicate experimental design.

Opinions

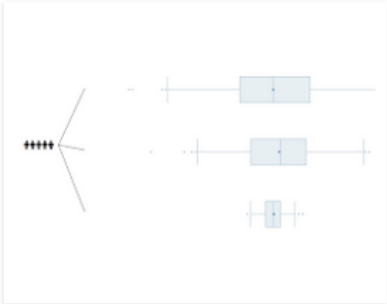

Which design did you like more? \*

	
<input type="radio"/> Black icons with arrow	<input type="radio"/> Coloured icons

Why did you choose that one? \*

Your answer \_\_\_\_\_

Which design did you like more? \*

	
<input type="radio"/> Black icons with arrow	<input type="radio"/> Coloured icons

Why did you choose that one? \*

Your answer \_\_\_\_\_

**Figure B.4:** Questions asking for the participant’s preference for indicating experimental design.

### B.1.1 Qualitative Feedback

#### Between-Subjects Design

Quotes from participants who preferred the coloured icons (8 out of 9 participants):

- "easier to tell visually"
- "I think the colors represent that, for example, yellow people are in the same group or they have the same function so you can clearly see that there are 3 groups in this case."
- "It's easily noticable with the color difference that there were 3 different groups"
- "Becuase if we use color it is posible to remove the arrows so the information representation become simpler and clearer"
- "More clear that its representing 3 different groups"
- "It is easier to distinguish the difference"
- "Less boring than the black ones"
- "Better to compare them"

Quotes from participants who preferred the black icons (1 out of 9 participants):

- "Easier to understand"

#### Within-Subjects Design

Quotes from participants who preferred the coloured icons (4 out of 9 participants):

- "easier to tell visually"

- “The color was at first confusing to me. It took me time to figure out that it represents all geoup took all the exams. Again for the black version it could appear that Only one group took all three exams. So i would choose color version but i think the design is bit confusing and could be better.”
- “Information is simpler than the black one”
- “the second one assign the same group with colours instead of arrows which is great!”

Quotes from participants who preferred the black icons with arrow (5 out of 9 participants):

- “The colored design makes it confusing whether the people are the same or that they are random for each exam. It’s not possible to know from the colored design if the same people are doing the different exams.”
- “Easier to understand which sample group was”
- “Clearer that it’s one group only”
- “It makes me understand that the three blockplots are related”
- “I was not sure about the second one (colored design)”

## B.2 Evaluation

### B.2.1 Scenario Used for Phase 1 of Evaluation

**Scenario** A pizza chain restaurant would like to boost its sales by investing in more advertisement. They have created two different commercials to be aired across Germany. The data shows the increase in pizza sales across all restaurant locations in the country.

- **Purpose of the Study:** Compare the increase in pizzas sold as a result of airing two different commercials.
- **X-Axis:** The x-axis is the increase in pizzas sold after airing the commercials. 0 means that the pizza sales did not increase.
- **Box Plot #1:** The data points in Box Plot #1 are the restaurants affected by Commercial #1.
- **Box Plot #2:** The data points in Box Plot #2 are the restaurants affected by Commercial #2.
- **Locking an Outlier:** One restaurant has poor customer service, and its pizza sales are not affected by the success of the commercial.
- **Setting a Boundary (Upperbound):** One restaurant has limited open hours, and can only sell a maximum of 30 pizzas.

### B.2.2 Changes to StatPlayground Following Phase 1 of Evaluation

#### Changes to the Assumption Checks

- Added tooltips to the homogeneity of variance component and the normality component, that would read the name of the assumption and whether it passed or failed. (e.g., "Homogeneity of Variance: Pass")
- Added pseudo-implementation for the normality test, so that it would fail if the histogram had more than one peak, or if the tails were very different.

#### Changes to the Fine Controls

- Fine controls no longer open automatically. Rather, the user has to double-click on a property to open the fine control menu

- Changed the icons to set the upperbounds and lowerbounds. Instead of using arrows, the icons were replaced with "[..." and "...]". This was meant to eliminate the problem of mistaking the upperbounds and lowerbounds properties as indicators or controls for moving the property
- The graphics used to represent the upper and lowerbounds were made smaller, so as not to distract the user
- The upper and lower bounds are not always shown, only when the property is selected

### Changes to the Results

- Instead of displaying the results above each boxplot, they are placed at the bottom as summary
- The results section was given a title "Results"
- Hide results for chosen test (since not implemented)
- Add placeholder "N/A" strings when there are not enough results to produce

### Addition of a Tutorial Page

A tutorial was made to replace the need for an instructor. It goes through all the features offered by StatPlayground. This tutorial is created as a separate page on the StatPlayground application, and is the first page the user sees upon opening the application. The tutorial is not interactive, but the user can easily switch back and forth between the tutorial page and the interactive page.

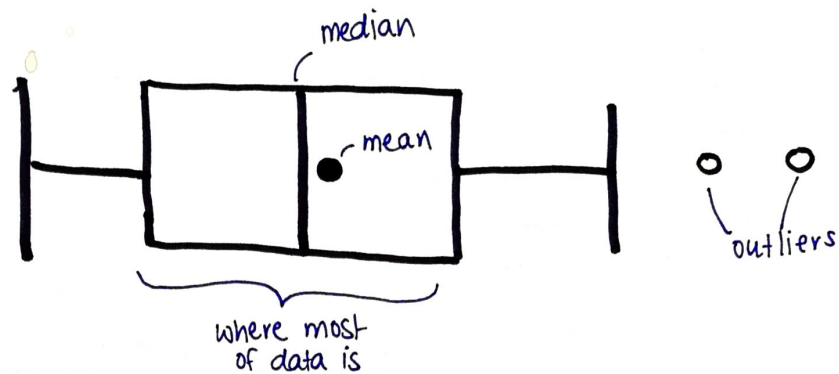
### Other Changes

- Make it configurable to show/hide the confidence intervals

- Show component for experimental design
- Remove feature to select the test type, since the feature was not implemented

### **B.2.3 Scenario and Tasks Given in Phase 2**





p-value : whether there is a difference  
Effect Size: how big the difference is

$p \leq 0.05$  reject  $H_0 \Rightarrow$  significant  
 $p > 0.05$  accept  $H_0 \Rightarrow$  they are the same



**Figure B.5:** An image of the paper handout presented to participants in Phase 2 of the evaluation. This page describes the properties of a boxplot, similar to that used in StatPlayground. The page also provides short summaries on the terms *effect size*, *p-value*, and [statistically] *significant*

## StatPlayground User Study

### January 2019

#### Scenario

Researchers at the RWTH are designing a new interface for Campus Office that is meant to reduce the amount of stress experienced by new users. One of the factors they choose to measure is the increase of heartrate from resting heart rate.

#### Setup

The X-axis represents the change of heart rate from resting heart rate. 0 means that there is no change from resting heart rate.

The first distribution (box plot) will represent the heart rates of the group that used the old design for Campus Office.

The second box plot will represent the heart rates of the group that used the new design for Campus Office.

Null Hypothesis: There is no difference in increased heart rate between the old design and the new design.

**Figure B.6:** An image of the paper handout presented to participants in Phase 2 of the evaluation. This page describes the example scenario used to provide context for the participant, and describes what components in StatPlayground describe the different variables in the scenario.

## Tasks

### Setup

1. Set the first box plot (old design users) so that the mean increase in heart rate is 40. This is our control group.
2. Add a new box plot to represent the users of the new design.

### Manipulate the Data

3. How might the dataset look if the new design did not affect the change in heart rate?
4. How might the dataset look if the new design reduces the amount of stress (i.e., minimizes the increase in heart rate) with a medium effect size?
5. Change the distributions so that the following assumptions are both met:
  - a. Homogeneity of Variance: The spread of all distributions are similar
  - b. Normality: Each set of data is normally distributed
6. Create an outlier to represent a user that drank espresso before using the new design. This user would have a higher heart rate than the rest of the group.
7. How would you manipulate this outlier, so that they would have the same heart rate regardless of how stressful the user interface is?
8. How would you manipulate this outlier, so that the heart rate will increase by a minimum of 55bpm, regardless of how stressful the user interface is?

**Figure B.7:** An image of the paper handout presented to participants in Phase 2 of the evaluation. This page contains the list of tasks that the user must complete.



# Bibliography

American Psychological Association et al. *Publication manual*. American Psychological Association Sixth Edition. Washington, DC, 1994.

Olivier Bau and Wendy E Mackay. Octopocus: a dynamic guide for learning gesture-based command sets. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*, pages 37–46. ACM, 2008.

Paul Cairns. Hci... not as it should be: inferential statistics in hci research. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1*, pages 195–201. British Computer Society, 2007.

Dane Coffey, Chi-Lun Lin, Arthur G Erdman, and Daniel F Keefe. Design by dragging: An interface for creative forward and inverse design with simulation ensembles. *IEEE transactions on visualization and computer graphics*, 19(12):2783–2791, 2013.

Adrian Furnham. Response bias, social desirability and dissimulation. *Personality and individual differences*, 7(3):385–400, 1986. doi: 10.1016/0191-8869(86)90014-0. URL [https://doi.org/10.1016/0191-8869\(86\)90014-0](https://doi.org/10.1016/0191-8869(86)90014-0).

Iddo Gal. Adults' statistical literacy: Meanings, components, responsibilities. *International statistical review*, 70(1):1–25, 2002. doi: 10.1111/j.1751-5823.2002.tb00336.x. URL <https://doi.org/10.1111/j.1751-5823.2002.tb00336.x>.

Joan Garfield. How students learn statistics. *International*

- Statistical Review/Revue Internationale de Statistique*, pages 25–34, 1995.
- Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.
- Brittany Kondo and Christopher Collins. Dimpvis: Exploring time-varying information visualizations by direct manipulation. *IEEE transactions on visualization and computer graphics*, 20(12):2003–2012, 2014.
- David M Lane and S Camille Peres. Interactive simulations in the teaching of statistics: Promise and pitfalls. In *Proceedings of the Seventh International Conference on Teaching Statistics*, 2006.
- Justin Matejka and George Fitzmaurice. Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1290–1294. ACM, 2017.
- Raymond S Nickerson. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, 5(2):241, 2000.
- Matthias Nielsen, Niklas Elmqvist, and Kaj Grønbaek. Scribble query: fluid touch brushing for multivariate data visualization. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*, pages 381–390. ACM, 2016.
- Don Norman. *The design of everyday things: Revised and expanded edition*. Constellation, 2013. ISBN 978-0-465-05065-9.
- Luis Carlos Silva-Ayçaguer, Patricio Suárez-Gil, and Ana Fernández-Somoano. The null hypothesis significance test in health sciences research (1995-2006): statistical analysis and interpretation. *BMC medical research methodology*, 10(1):44, 2010.
- Mads Soegaard. Gestalt principles of form perception. <https://www.>

interaction-design.org/literature/book/the-glossary-of-human-computer-interaction/gestalt-principles-of-form-perception, 2015. Accessed: 2019-01-31.

Krishna Subramanian and Jan Borchers. Statplayground: Exploring statistics through visualizations. In *CHI '17: Extended Abstracts of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 401–404, New York, NY, USA, May 2017. ACM. ISBN 978-1-4503-4656-6. doi: 10.1145/3027063.3052970. URL <http://doi.acm.org/10.1145/3027063.3052970>.

Harold Thimbleby. Supporting diverse hci research. In *Proceedings BCS HCI Conference*, volume 2, pages 125–128. Citeseer, 2004.

Jessica Utts. What educated citizens should know about statistics and probability. *The American Statistician*, 57(2): 74–79, 2003.

Romain Vuillemot and Charles Perin. Investigating the direct manipulation of ranking tables for time navigation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2703–2706. ACM, 2015.

Friedrich Wulf. Beiträge zur psychologie der gestalt. *Psychological Research*, 1(1):333–373, 1922.

Alain F Zuur, Elena N Ieno, and Chris S Elphick. A protocol for data exploration to avoid common statistical problems. *Methods in ecology and evolution*, 1(1):3–14, 2010. doi: 10.1111/j.2041-210X.2009.00001.x. URL <https://doi.org/10.1111/j.2041-210X.2009.00001.x>.





# Index

abbrv, *see* abbreviation  
assumption checks, 29–30

Cohen's  $d$ , 15, 31  
colour, 19, 20, 31, 40, 42, 44  
confidence interval, 26

design approach, 13–24  
Design by Dragging, 8–9  
design principles, 14  
DIA cycle, 13  
DimpVis, 10  
discussion, 43–44

effect size, 19, 20, 31, 32, 39, 43, 46  
evaluation, 33–44, 57–64  
experimental design, 26, 51–57

fine control, 27–29  
foreshadowing, 10, 18, 31  
future work, 47

gestalt laws, 43–44

hints, 10, 18, 31  
homogeneity of variances, 29

interaction design, 25–32  
introduction, 1–5  
iterative design, 13

learning effect, 34, 38  
limitations, 42–43

normality of distributions, 29

OctoPocus, 10  
outline, 5

p-value, 31, 32, 46

participants, 34, 38

questionnaire, 20–22, 51–57

research questions, 4

response bias, 43

results, 31–32

summary, 45–46

t-test, 14

think-aloud protocol, 44

tooltip, 30, 58

tutorial, 36, 42, 59

user study, 16–20, 33–42

