Statsplorer: Guiding Novices in Statistical Analysis

Chat Wacharamanotham **Krishna Subramanian** **Sarah Theres Völkel** **Jan Borchers**

RWTH Aachen University
52062 Aachen, Germany
{chat, krishna, borchers}@cs.rwth-aachen.de, sarah.voelkel@rwth-aachen.de

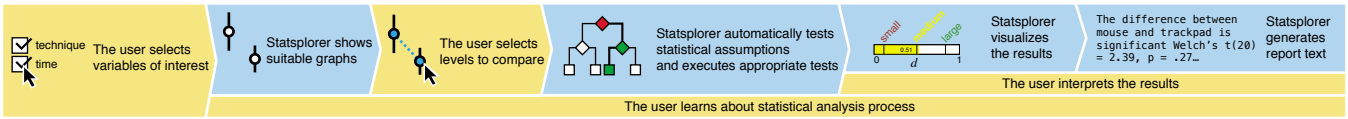

Figure 1. Basic process of statistical analysis with Statsplorer

ABSTRACT

Each step of statistical analysis requires researchers to make decisions based on both statistical knowledge and the knowledge of their own data. For novice analysts, this is cognitively demanding and can lead to mistakes and misinterpretations of the results. We present *Statsplorer*, a software that helps novices learn and perform inferential statistical tests. It lets the user kick-start data analysis from their research questions. Statsplorer automatically tests necessary statistical assumptions and uses visualizations to guide the user in both selecting statistical tests and interpreting the results. We compared Statsplorer with a statistics lecture and investigated how Statsplorer prepares novices for learning statistics in an AB/BA crossover experiment. The results indicates that using Statsplorer prior to the lecture leads to significantly better test scores in understanding statistical assumptions and choosing appropriate statistical tests. Statsplorer is open-source and is available online at: http://hci.rwth-aachen.de/statsplorer.

Author Keywords

Inferential statistics; data analysis; data visualization

ACM Classification Keywords

H.5.2 User Interfaces: Graphical user interfaces (GUI);
G.3 Probability and Statistics: Statistical software.

INTRODUCTION

Problems in using inferential statistical analysis for research are prevalent in medicine [1], biology [39], psychology [8] and HCI [6, 13, 18, 20]. A potential explanation is that, despite widespread availability of educational resources (e.g., [22, 31, 36]), learning statistics is difficult and takes time [15, 16, 38]. Hence, novice analysts resort to just-in-time learning, an approach that is prevalent in software development and general trouble-shooting [5, 23]. An indicator of this practice is the proliferation of online forums for statistical problems[1].

During statistical analysis, novices can be easily overwhelmed by the number of decisions to make and relevant information to consider. Below is a typical example of analysis procedure from textbooks (e.g., [14, 22]):

The analyst needs appropriate graph (e.g., histogram or boxplot) to learn about characteristics of her data (e.g., shapes of the distribution, outliers). She also needs to test statistical assumptions (e.g., normality, homogeneity of variance, independence). Based on the characteristics, assumptions, and how the data was acquired (e.g., experimental design, the scale of the measured data), she needs to decide upon the appropriate statistical test (e.g., t-test, ANOVA, RM-ANOVA, Kruskal-Wallis) and follow-up test (e.g., Tukey test, pairwise t-test with a proper adjustment). Interpretation and write-up also require careful consideration of the above information while conforming to statistical reporting standards (e.g., APA Style Guide [2]). Additionally, to perform the aforementioned tasks the novice analyst needs to be familiar with her statistical analysis software. Keeping track of all this information, while trying to understand various statistical analysis procedures, can be challenging, especially for novices [12].

To help novice analysts learn and perform statistical analysis in an explorative manner, we developed and evaluated Statsplorer, a software that guides them through different steps in statistical analysis. Statsplorer provides appropriate data visualizations and assists in making decisions (Fig. 1). This paper makes the following contributions:

- Statsplorer, an open-source software that helps novices to perform and learn statistical analysis
- An evaluation that shows how Statsplorer benefits novices

In the next section, we review existing software and research works that support statistical analysis in empirical research.

CHI 2015, April 18–23, 2015, Seoul, Republic of Korea.

ACM 978-1-4503-3145-6/15/04 ...$15.00.
http://dx.doi.org/10.1145/2702123.2702347

[1] E.g., stats.stackexchange.com and reddit.com/r/statistics

RELATED WORK

Data collection and analysis in empirical research consist of several steps. We first describe ecology of existing software that support these steps to indicate Statsplorer's niche. Then, we review the learning principles pertinent to statistics that are used in Statsplorer.

Software Support in Empirical Research

Tools for performing preliminary steps in empirical research, such as designing experiments and cleaning data, already exist in the literature. For designing experiments, which is usually the first step in empirical research, researchers can use Touchstone [25] to explore experimental designs and generate sequences of conditions with proper counterbalancing. Then, during the experiment, Touchstone signals the software used in the experiment to administer these conditions in the experiment and to collect the data. After the experiment, researchers can use Wrangler [19] to create a script to preprocess the data into a format suitable for statistical software.

Subsequently, researchers analyze the data to answer research questions that fall into three categories: base rate, correlation, or differences [27] For base rate questions, tools such as Kinetica [32] and TouchViz [11] show that interacting directly with data through graphs helps the users understand data better and answer questions on descriptive statistics faster and with less errors. For correlation questions, EvoGraphDice uses an interactive scatterplot matrix generated by evolutionary algorithm to help users discover complex relationships in multi-dimensional data [4]. For research questions involving differences, we discuss the existing tools below.

Software that Help Users in Statistical Analysis

Mainstream software, e.g., SPSS, SAS, JMP, R, and Tableau require the analyst to know, a priori, (1) what graph to plot and what statistical test to use, (2) what information is relevant from the many tables and graphs that are generated in statistical reports, and (3) how to interpret them.

Software designed for novices[2] determine the appropriate statistical test from the data type of selected variables. (This approach was pioneered by VisTa [37].) However, these software perform predefined statistical tests and implicitly make statistical assumptions without users' awareness. AdviseStat[3] generates statistical reports that additionaly provide theoretical information of the statistical results. However, AdviseStat only visualizes data as results, instead of using data visualizations to help users make decisions about statistical tests.

VisTa [37] represents each analysis workflow as a directed cyclic graph. Users can experiment with the workflow by activating or deactivating each step of the analysis. However, the analyst requires expertise to construct the workflow. Also, to follow the workflow, the analyst requires knowledge to decide on statistical procedures and to interpret the results.

Illmo [26] guides users to perform Thurstone modeling analysis. Similar to VisTa, Illmo uses block diagrams to represent steps in the modeling process and visualizes the results. Using Illmo to interpret and make decisions on the data analysis procedure still requires knowledge of the modeling.

Statsplorer aims to help users perform statistical analyses and to help them understand the decisions made during the analyses. To achieve this, Statsplorer incorporates several principles from research in statistical education.

Principles in Statistical Education

Dufresne et al. showed that by constraining problem-solving to explicit steps, each of which requires a decision among a few choices, students are more aware of the principles and procedures in problem-solving [12]. Garfield conducted a meta-analysis to derive a set of principles for learning statistics in 1995 and re-validated them in 2007 [15, 16]. In Statsplorer, we used the following principles, which Garfield suggested to statistics instructors: (1) let students construct knowledge based on their prior knowledge, (2) allow guesses and predictions to be confronted with actual results through real-time feedback, and (3) use technology to visualize and explore data and statistical models [16]. Lovett and Greenhouse derived similar principles from cognitive theory and additionally suggested to lessen mental load of students by showing only the necessary information [24].

Statsplorer incorporated these principles to support novice users in statistical analysis. Statsplorer guides users through a constrained analysis path and uses interactive visualizations to help them identify problems in the data and interpret the results. To identify the initial set of statistical tests that Statsplorer should support in its initial version, we surveyed HCI literature in the manner described below.

STATISTICAL ANALYSIS IN CHI

In 2007, Cairns surveyed statistical analysis usage and problems from four issues of three HCI journals [6]. Out of 80 papers, 41 use statistics, of which half use null hypothesis significance testing (F and t tests).

To verify whether the findings still hold true, we surveyed statistical methods used in the Paper and Notes venue of CHI 2014 (single annotator). More than half of the papers (255 out of 465 papers, 54.84%) contains statistical analysis more advanced than descriptive statistics (e.g., percentage or means). Many of these papers use multiple tests. Out of 255 papers, 122 use ANOVA (47% of those that use statistics). Other central tendency comparisons (e.g., t-tests, Wilcoxon rank sum tests) are used in 53 papers (20%). Chi-square test of relationships are used in 22 papers (8%). 65 papers (25.49%) used regression analysis. Similar to Cairns' survey, our survey found that null-hypothesis significance testing (e.g., ANOVA, t-tests, Wilcoxon rank sum test) is still widely used. This proportion suggests that the choices of statistical analysis procedures have not changed much over the last seven years.

Although several alternative analysis methods have been proposed since the original survey [10, 20, 26], none of them are widely used. Therefore, we chose to focus on the null-hypothesis significance testing approach in the initial version.

[2] E.g., wizardmac.com, statwing.com

[3] adviseanalytics.com (discontinued as of December 2014)
Demo video: http://youtu.be/kMfwC1f7LNY

INTERACTION DESIGN

First, we briefly describe how a user might use Statsplorer to analyze data[4]. To analyze data in Statsplorer, the user selects variables of interest. Statsplorer, then, automatically selects a graph, based on the number of variables selected and their roles, that is likely to give interesting insights . This graph is a starting point for further analysis, and always stay on the screen until the user changes variable selection. Such graphs are interactive. For example, the user can hover the mouse cursor over the graph to see descriptive statistics, e.g., a mean and its 95% confidence interval. To compare means, the user switches to a comparison mode and directly selects the means of distributions on the graph. Then, Statsplorer automatically performs statistical assumption tests and selects an appropriate test to compare the mean differences. In case assumptions are violated, Statsplorer determines whether data transformation or alternative tests are possible and suggests possible actions. After executing the test, the results are shown in an interactive graph to help the user interpret effect size and confidence interval of the difference. Additionally, Statsplorer generates a statistical report text that can be readily used, e.g., in a paper.

Design Principles

To help novices perform and understand statistical analysis, we designed Statsplorer based on two principles.

1. *Statsplorer is visualization-driven:* Statsplorer emphasizes interaction with data visualization in order to initiate *any* analysis task, to show the statistical analysis process, and to guide interpretation of results. Unlike mainstream statistical software, e.g., SPSS, graphs are not optional in Statsplorer. Statsplorer keeps the data in context throughout the entire analysis in order to reduce cognitive load of the user, and to make them aware of problems that may stem from the data characteristics.

2. *Statsplorer guides the user through explicit, narrow, and deep decision trees:* Statsplorer explicitly shows decisions made during statistical procedures to the user. This makes the user aware of the statistical analysis procedure and the assumptions that are made. Statsplorer also recommends default course of actions and allows the users to select alternative tests according to their knowledge of the characteristics of the data.

In the following sections, we describe how Statsplorer applies these principles to address the four major statistical problems in HCI [6]: *ignoring statistical assumptions, inappropriate testing, over-testing,* and *incorrect or non-standard reporting*.

Automatic Selection of Graphs and Statistical Tests

Statsplorer automatically determines the appropriate data visualization based on the number of selected variables and their roles. Statsplorer uses only four simple, yet powerful, data visualizations to avoid confusing inexperienced users. Table 1 shows the purpose of each visualization. Descriptive statistics are shown when the user hovers with her mouse over the visualized data.

[4]A walk-through is also available in the video figure.

Table 1. Four data visualizations used in Statsplorer

Graph	Purpose	Triggerred on selection
Histogram	• shows shapes of the distribution • shows potential bimodal distribution	single variable
Boxplot	• shows outliers • shows central tendencies and spread	1 dependent variable and 1–2 independent variables
Scatterplot	• show correlations	2 dependent variable and 0–1 independent variable
Matrix of scatterplot	• show correlations of >2 variables	three or more variables

When the user selects means to compare, Statsplorer automatically tests the normality and homogeneity of variance assumptions. Based on the results of these tests and the data type of variables, Statsplorer executes an appropriate statistical test. The decision steps that govern the test selection are shown in a decision tree at the top of the screen. Among the tests that are appropriate in the given situation, Statsplorer always chooses the test that maximizes statistical power[5]. However, the user can expand the decision tree and select alternative tests for exploration. Also, sometimes, the user may need to override Statsplorer's decision based on her knowledge of the data characteristics. For example, to assess homogeneity of variance, Statsplorer uses Levene's test, which may be too sensitive to a small deviation from normality [14], especially in a large dataset. Therefore, it makes sense for the user to select alternative tests based on the data characteristics, e.g., data distribution visualized in the diagnostic plot (Fig. 2).

Upon violation of any of the assumptions, Statsplorer notifies the user by appropriately animating the main visualization. The user can click on each assumption to show the diagnostic plot and statistics concerning the assumption. For the normality assumption, we used histograms, with an improved method for bin size calculation to provide better estimates [35], to show the distribution shape. For homogeneity of variance, we visualize the confidence interval of SDs calculated from a χ^2 distribution. Statsplorer also tests, in background, whether any data transformations ($log(x)$, $\sqrt{x}$, $\sqrt[3]{x}$, $\frac{1}{x}$) satisfy these assumptions and suggests suitable actions to the user.

In the test results, Statsplorer visualizes the standardized effect sizes in a progress bar with legends indicating how to interpret them. The mean differences and their confidence intervals are also shown when the user hovers over the data she has compared. For interaction effects, Statsplorer constrains users to focus on higher-order effects before interpreting the main effects. Interactive interaction plots allow users to change the levels of each independent variable to compare their effect with other independent variables.

Over-testing Detection and Correction

Statsplorer tracks all comparisons that the user performs. This allows Statsplorer to detect two possible types of over-testing: (1) when multiple t-tests are used instead of an ANOVA, and (2) when multiple one-way ANOVAs are used instead of a factorial ANOVA (likewise for multiple 2-way ANOVAs instead of a 3-way ANOVA). Nevertheless, whether a sequence of action is over-testing or not depends on the

[5]See supplements for full details

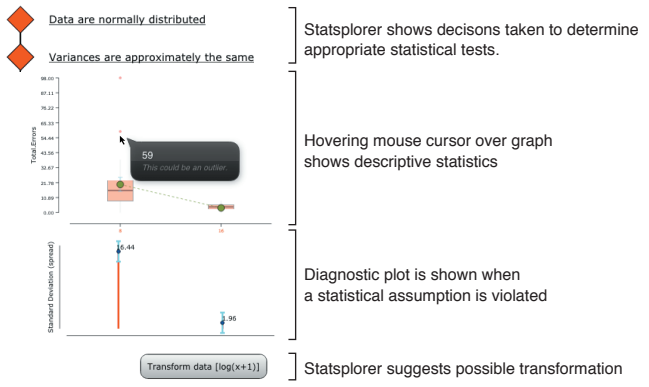

Figure 2. Statsplorer screen, showing a statistical assumption violation

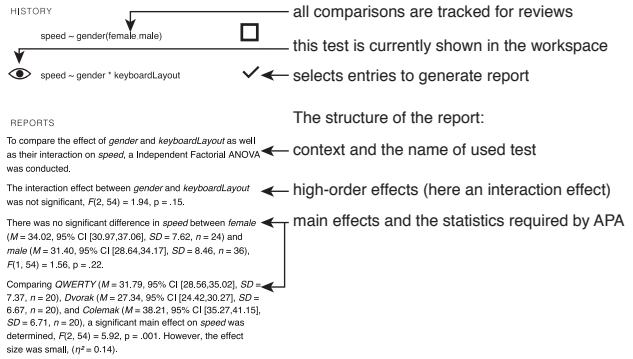

Figure 3. The user can select a set of test to genrate textual report.

user's research questions. Therefore, when a possible over-testing is detected, Statsplorer asks the user to clarify research questions and perform higher-order test when appropriate.

Generating Statistical Reports

Statsplorer generates statistical reports of the tests performed in a format conforming with the APA standard [2]. Statsplorer uses short sentences and explicit causal connectors to improve comprehensibility for novices [3, 28]. The sentences are ordered according to proper statistical interpretations, e.g., describing the interaction effects prior to the main effects as shown in Figure 3.

SYSTEM ARCHITECTURE AND INTEGRATION

Statsplorer uses a client-server architecture[5]. The statistical computation back-end of Statsplorer is written in R and communicates with the client through OpenCPU [30]. On the client side, Statsplorer is implemented in Javascript and runs in standard web browsers.

We designed Statsplorer to be a platform for prototyping and testing statistical analysis UIs. Since the only requirement for Statsplorer is a web browser, it is possible to test the prototyped UIs online, e.g., using mechanical turk services. Statsplorer can be installed locally or on a web server that supports R or on a free OpenCPU service[6].

[6]https://www.opencpu.org

USER STUDY: EFFECT OF STATSPLORER IN LEARNING

Although it is possible to compare Statsplorer with other software tools, using such tools requires users to be knowledgable beyond the novice level. Such users are not our main target group. Therefore, we compared Statsplorer with an introductory statistics lecture. This allows us to control the exposure duration, match learning objectives, and recruit novices with adequate motivation to learn. Such recruitment method is also recommended by the literature [26, 37].

Since Statsplorer makes statistical analysis process more visible, we additionally investigated the extent to which this helps novices to *prepare for future learning* (construct their own theories about the phenomena before contrasting it with the knowledge from the expert). Preparing students to learn is shown to allow students to transfer knowledge better than *tell-and-practice* (e.g., doing exercises after a lecture) in descriptive statistics [34] and neuroscience [33]. Therefore, we used an AB/BA crossover experimental design in the same way as [33]. Participants in Group A used Statsplorer before the lecture, and vice versa in Group B (Fig. 4). This design allows us to compare the effect of Statsplorer vs. the lecture (in the mid-test) and the effect of sequence (in the post-test).

Therefore, we have two hypotheses:

H1: In the mid-test, participants who use Statsplorer (Group A) will score better than those who attended the lecture (Group B).

H2: In the post-test, participants who use Statsplorer prior to the lecture (Group A) will score better than those who use Statsplorer after the lecture (Group B).

As in [33], we triangulate quantitative and qualitative evidences in the manner described next.

Procedure Overview

Figure 4 provides an overview of the procedure of this study. Details on the stimuli (the tests, task sheet, and lecture) are explained in the Stimuli Design section. To minimize bias from students in their behavior, the entire procedure was conducted by one of the authors, who is a college student and not a part of the team that organizes the lecture. We used the constructive interaction method in Statsplorer sessions. We capture users' spontaneous comments and discussion during the sessions to analyze their insights [29, 33].

Pre-test: First, participants were asked to fill out the pre-test. Based on the score of the pre-test, we split them into two groups, and then into pairs for constructive interaction. We ensured made both groups equivalent in their initial knowledge by balancing participants with similar test score. We also matched the participants' level of knowledge within each pair. To prevent bias, we did not inform participants about the rationale of grouping.

Statsplorer and mid-test (Group A): Then, each pair of the participants in Group A used Statsplorer in a lab. First, they were introduced to Statsplorer with a short video walk-through. This video shows only the interaction flow without explaining any procedures in statistics. Then, the participants

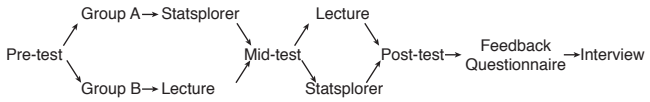

Figure 4. To compare the effect of the system and that of the sequence, we used an AB/BA experiment.

used Statsplorer to analyze datasets to answer statistical questions on a given task sheet. An earlier version of Statsplorer was used in this study[5]. The report generation was not on the same screen as the analysis. There was no decision tree, and it was not possible to select alternative statistical tests. We limited the overall time to 50 minutes. For those who finished earlier, we encouraged them to use the system to learn more about statistics if they wanted to. An experimenter stayed in the room to observe and take field notes as well as to help when the participants were confused. Nevertheless, the participants were asked not to pose any questions related to statistical procedures. We videotaped the interaction and took field notes for later analysis. The recorded videos are annotated in order to analyze the number of utterances that the participant generates.

After the session, both participants took the mid-test. The user study sessions for Group A were conducted within 8 days prior to the lecture.

Lecture, post-test (Group A), and mid-test (Group B): After this, all participants attended the lecture. Right after the lecture, participants were asked to fill in the mid-test (for Group B) or the post-test (for Group A).

Retrospective interview: Afterwards, we asked participants in Group A to return, in assigned pairs, for a retrospective interview to elicit their learning progress and overall experience. Participants also filled in a written questionnaire evaluating their opinion on both Statsplorer and the lecture in this step. After the interview, the study for Group A was complete.

Statsplorer and post-test (Group B): Within a week after the lecture, each pair of participants in Group B used Statsplorer, filled in the post-test, and participated in retrospective interview. All of these steps were conducted in a single session in the lab. After the study was concluded, students were informed that they could access Statsplorer online for general use and learning.

Participants

Since the study involves multiple activities spanning over several days, we carefully recruited participants with similar background knowledge in statistics and ensured they were motivated to participate in the study. Our participants were students in a HCI research course taught by two of the authors[7]. In this course, students learn to analyze and critique research publications in HCI. In a lecture in this course, students are to learn the basic knowledge needed to understand and interpret statistics in HCI research papers. Through lectures and assignments prior to the test period, students are already familiar with basic concepts of empirical research methods such as experimental design, independent and dependent variables, and null hypotheses. Therefore, the lecture used in this study focuses on applying these concepts in statistical context to understand the results. We ran this user study one week before (Group A) and one week after the lecture (Group B) to minimize confounding effects from students' learning and revisioning style.

[7]Current Topics in Media Computing and HCI. Details available at http://hci.rwth-aachen.de/cthci_ss2014

Table 2. We balanced user's background knowledge between groups.

		Group A	Group B
n: Number of users in group		16 (4 female)	18 (5 female)
Age *M* ± *SD*		24.45 ± 1.65 years	24.45 ± 2.74 years
Pre-test score (*M*, 95% *CI*)		4.88 [2.35, 7.41] %	4.97 [2.32, 7.63] %
n by pre-test score	Low (0–1)	7	6
	Medium (2–4)	5	9
	High (4.5 – 10.5)	4	3
n by program of study	CS and related (M.Sc.)	11	14
	Technical Communication (M.Sc.)	5	4
n by previous statistics background	School	11	13
	1 university course	11	12
	> 1 university courses	3	6
	Used in seminar, thesis	3	6

(CS: Computer Science and related fields)

As an incentive, participants were awarded 3% of the course score. Students who did not participate in the study could do an extra exercise in statistics to achieve the same score. Since the students were from the HCI research course, we also debriefed the experiment in a practice session of the class after the conclusion of the study.

After screening the participants with the score from the pre-test to ensure similar background knowledge in statistics, we ended up with 34 participants. The participants' background in each group is shown in Table 2. Since all participants were college students, all of them took classes in basic probability and statistics either in school or the university. Although nine of them used statistics in their thesis or seminars before, their test results did not differ from other participants. The uneven number of participants is a result from issues in scheduling with the participants. Nevertheless, the average pre-test scores of both groups were similar.

Stimuli Design

Data analysis task: We designed four synthetic datasets, each from a distinct hypothetical experiment in the context of HCI and psychology. To ensure that all participants interacted with Statsplorer to a similar extent, we gave users a list of research questions in an exercise sheet format. These questions can be answered by checking descriptive statistics and performing and interpreting results of t-tests, one-way ANOVAs, post hoc tests, one-way repeated-measure ANOVA, and two-way ANOVA in this order[5]. We did not give the participants any instruction of the name of the tests to be executed. The sequence of the task makes it possible for the participants to face the over-testing problem by performing multiple t-tests on different pairs of levels of an independent variable to answer the research question. On average, participants took 45 minutes to solve these questions.

Lecture: The lecture lasted 90 minutes. It reviewed descriptive statistics concepts (e.g., central tendency and confidence interval), introduced effect sizes (e.g., Cohen's d) and explained the process of null hypothesis significance testing, p-value and its interpretation. The lecture covered assumptions in parametric statistics with focus on normality and homogeneity of variance. Lastly, the lecture summarized how to apply these concepts to select statistical tests as a decision tree. The lecture was designed based on principles of statistics instruction design and statistics learning [16, 24]: The lecturer used simulations from [7] to show the relationship between mean, CI, and p-value. The lecturer also engaged the students by allowing them to work with a partner in small in-class exercises and question and answering.

Measurements and Data Analysis

Tests for assessing statistical knowledge: Existing instruments in statistics education (e.g., [17]) focus on fundamental concepts, which is not the knowledge that Statsplorer aims to communicate. We developed a set of questions[5] that focus on the knowledge that the users could learn from Statsplorer. These questions assess respondents' knowledge based on basic knowledge (6 points) and the four problems identified by Cairns [6]: assumptions (7), appropriate test selection (14), over-testing (2), and reporting (23.5). Table 3 shows how these questions address different dimensions of knowledge according to the revised Bloom's taxonomy [21]. These questions are checked by the course instructor and an external statistician. Since the spread of score is neither complete nor balanced between cells, in our analysis we ignore the cells that have only one point to avoid over-sensitivity.

From this set of questions, we created three isomorphic tests to be used as pre-, mid-, and post-test: they use the same core concepts but with different numeric values or narrative examples used in the questions, order of choices, and negations. All questions and answer keys are designed by one of the authors and checked by another author. We administer these tests via an online form and on paper, depending on the preference of the participants. Participants were asked to fill in the test without consulting any aid in statistics.

To compare the differences of the test scores, we consider score from each type of Cairn's four problems and each cells in Table 3 as a dependent variable. We used the test depending on whether the distribution satisfies the normality assumption (determined by Shapiro-Wilk test) and homogeneity of variance assumption (determined by Levene's test). If both assumptions are satisfied, we use unpaired t-tests. If the homogeneity of variance is not satisfied, we use Welch's t-test. In both cases, we report effect size with Cohen's d (using pooled within-groups SD; small: 0.2, medium: 0.5, large: 0.8). When assumptions are not satisfied, we use Mann-Whitney-U tests and effect size r (small: 0.1, medium: 0.3, large: 0.5). The differences are deemed statistically significant when $p < .05$. We use 95% confidence interval for error bars in graphs and CIs in the report.

Video and screen recording: We recorded the screen throughout the Statsplorer user study session. Additionally, we videotaped the session from the back of the participants to capture the discussion that the participants made, e.g., by pointing the finger to refer to specific parts of the screen. We coded the participants' utterances according to the categories in [33]. This coding scheme were chosen a priori. Table 4 shows example quotes from each category.

Table 3. Test questions cover a subset of knowledge dimensions

	Cognitive process					
Knowledge	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual	1, 3, 3, –, 12.5 = 19.5 points	1, –, –, –, – = 1 point				
Conceptual		3, –, 2, 1, – = 6 points		1, –, –, –, – = 1 point		
Procedural		–, 1, 3, –, – = 4 points	–, –, –, –, 1 = 1 point		–, 3, 6, 1, – = 10 points	–, –, –, –, 10 = 10 points
Legend:	basic knowledge, assumptions, test selection, over-testing, reporting = total score					

Table 4. The utterances coding categories with examples from our users

Category and explanation	Example
Simple observation	"…and not we have better effect size than we have earlier"
Prediction	"I took [the data] without [transformation]. So [that] when they used Welch's ANOVA."
Confrontation between a prediction and a result	"Let me [check the help on the ANOVA]", "So we go back and then we check? We transform it"
Definition of a rule or principle	"Yeah, we have ANOVA here, so that the effect size was Eta-squared"

Feedback questionnaire: The feedback questionnaire[5] used right before the retrospective interview was designed to assess overall experience of both Statsplorer and lecture. Additionally, we asked questions from Technology Acceptance Model [9] to evaluate perceived usefulness, usability, enjoyment, and perception of time during Statsplorer use.

Retrospective interview: We conducted semi-structured interviews. Our interview protocol[5] comprises of five parts: First, we asked participants to evaluate Statsplorer and the lecture with open-ended questions. Then, participants evaluated the interaction and the role of Statsplorer and the lecture in their statistics learning. In the third step, the experimenter presented research scenarios[5] as probes to assess their understanding in statistics. These probes are designed to cover all of the four statistics problems identified by Cairns [6]. In the next part, the participants were asked to imagine themselves as a teacher for statistics and were encouraged to suggest possible improvements for both Statsplorer and the lecture. In the last part, participants provided additional comments on both Statsplorer and the lecture. We recorded audio and the experimenter wrote down field notes during the interview session. The qualitative data from the interview is analyzed using the Grounded Theory method. One experimenter coded the data and iteratively categorized the results.

Results

Statistical Knowledge Tests

The results are shown in Figure 5. We provides our interpretations based on CIs, which allows us to discuss magnitude of the effects [7]. Some significant results ($p < .05$) are considered weak when the CI almost touches zero.

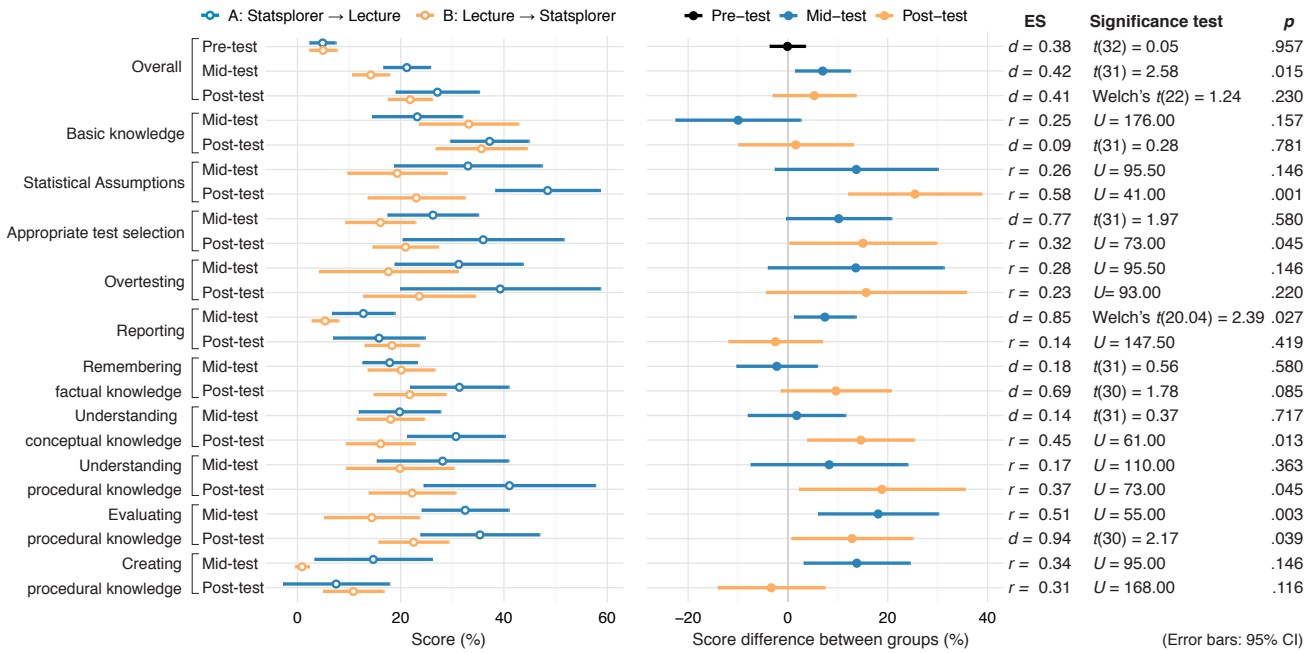

Figure 5. As most of the score differences are higher than zero, the result generally favors Statsplorer over the lecture. The result also indicate benefits of using Statsplorer before the lecture to prepare for learning, although the evidence was not as strong.

Overall: In the pre-test, Group A scored 6.96% higher than Group B with 95% CI of [1.47%, 12.46%]. This result suggests that using Statsplorer is more effective than attending the lecture. In the post-test, although Group A outperformed Group B by 5.28% [–3.00%, 13.59%], the CI of the difference crosses zero, providing only a weak evidence for the effect of the sequence. The CI of score in Group A is twice the size of that in Group B. This is probably because Group A retains less knowledge, especially in the questions in the creating procedural knowledge dimension.

Scores by Cairn's four problems: For the questions on statistical assumptions, the mid-test score of Group A surpassed Group B by 13.71% [–2.60%, 30.04%]. Although the difference is large, it is only a weak evidence because the CI slightly crosses zero. However, Group A scored better than Group B in the post-test by 25.45% [12.10%, 38.79%]. The range of difference is far from zero, providing a strong evidence for the effect of the sequence.

For the appropriate test selection questions, Group A performed better in the mid-test by 10.17% [–0.40%, 20.73%] and in the post-test by 15.05% [0.33%, 29.77%]. Although the estimates of both results favor our hypotheses, both evidences are weak.

Similarly, for the over-testing questions, Group A scored considerably better on average in both the mid-test (13.60% [–4.02%, 31.23%]) and the post-test (15.67% [–4.36%, 35.71%]). However, both of these results are only weak evidences (CIs cross zero). Since there are only two points for the over-testing questions, we have low statistical power for this category than the other.

For the reporting questions, Group A surpassed Group B in the mid-test by 7.4% [1.21%, 13.59%]. In the post-test, Group B was marginally better by 2.52% [–12.00%, 6.87%].

Scores by dimensions of knowledge: For the questions that test the understanding of conceptual knowledge, Group A was on par with Group B in the mid-test (1.75% [–8.0%, 11.53%]), but they outperformed Group B in the post-test by 14.60% [3.87%, 25.33%]. This shows a strong sequence effect despite the absence of the effect from just Statsplorer itself over the lecture.

For the questions that test the understanding of procedural knowledge, the difference in the mid-test has a large CI that crosses zero (8.27% [–7.4%, 23.99%]). But Group A clearly outperformed Group B in the post-test by 18.85% [2.24%, 35.46%], which supports the effect of the sequence.

For the questions that requires evaluating procedural knowledge, Group A clearly surpassed Group B in both the mid-test (18.09% [6.06%, 30.12%]) and the post-test (12.86% [0.73%, 24.99%]). This strongly supports both the effect of the system and of the sequence.

For the questions that require creating procedural knowledge, although Group A scored 13.81 % [3.18%, 24.43%] better in the mid-test. Group B was slightly better in the post-test (3.33% [–7.38%, 14.00%]). Even though the effect of the system is pronounced, the effect of the sequence is not.

Retrospective Interview and Field Notes

Learning through experimentation: Participants from both groups indicated that Statsplorer "allows … experimenting and … going into depth and thinking about why a specific test is chosen at a time." (n_A = 10/16 vs. n_B = 13/18). Participants suggested Statsplorer could act as "a starting point"

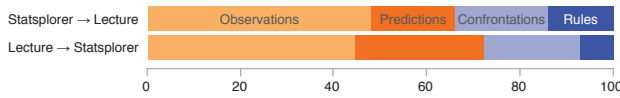

Figure 6. Average distribution of participant's utterances classified by categories. The difference between groups were not clear cut.

prior to the lecture (12 vs. 12). However, 10 participants from Group A mentioned that they aimed for "the right solution" rather than to learn the analysis process. (Only three from Group B mentioned this.)

Addressing the four problems: Participants praised that assumption testing is highly visible "[as a] check list" (13 vs. 9), and the reporting function was "useful" (4 vs. 7). The participants used the reporting function to "help understand the results [of the tests]" (7 vs. 11) and to "get the standard of reporting" (11 vs. 13). We observed that all groups skipped the visualization and results summary and read the report to understand the results.

Many participants "couldn't understand which test should be used at which point of time [because it was] not that visible" (11 vs. 11). They wanted more freedom and feedback in choosing statistical tests: "I couldn't say like I want to run this statistical test on it and then it would tell me no you can't because of . . . " (8 vs. 3). Similarly, participants "didn't really get the [process] of over-testing from [Statsplorer]" (5 vs. 9). Both problems were attributed to insufficient explanation of the decision process (5 vs. 0).

Exploration Utterances

Figure 6 shows the percentage of utterances in each category. Despite a tendency for Group A to make more observations and identify more rules, the total number and type of utterances greatly differs among the pairs of participants. On average, users in Group A uttered more rules than Group B in choosing appropriate tests (M_A = 4 rules, M_B = 2) and in statistical assumptions (M_A = 2, M_B = 1). Our users did not utter any rules about reporting or over-testing.

Feedback Questionnaire

Figure 7 shows the results of the feedback questionnaire. The results are positive across all dimensions except focused immersion. Although the participants rated high in temporal disassociation questions, they did not feel absorbed in using Statsplorer. This could be a side-effect from the constructive interaction method, in which participants interact with their teammate during Statsplorer use. Group B slightly rated Statsplorer more usable, probably because of their background knowledge from the lecture. The difference between groups in other ratings are neglegible.

Discussion

Statsplorer effectiveness: The users in both groups started with the similar knowledge level. After 45 min with Statsplorer, Group A performed better in the mid-test than those in Group B (90 min of the lecture). Group A rated Statsplorer useful, usable, and enjoyable. This shows that the users were able to perform statistical analyses and learn about it with Statsplorer, even without much preliminary knowledge. From

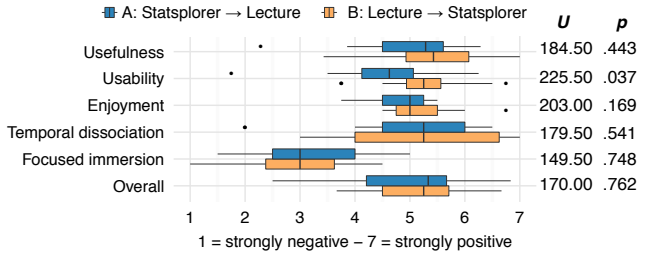

Figure 7. Regardless of the sequence, Statsplorer are rated positively in all dimensions except in the focused immersion.

the interview, we found that this is because the users were able to experiment and start using Statsplorer without requiring extensive knowledge in statistics. These results **support H1**. We now look into more details on the effectiveness of Statsplorer in solving the Cairn's four problems.

Although Group A scored better across all questions, the effect was most pronounced in the report questions. The users mentioned that they used the reports to understand the results.

Nevertheless, the evidence supporting other Cairn's problems are weaker. There was not much difference in the over-testing questions in both tests, probably because the given stimuli allows this feature to appear only once. The interview indicated that the explanation from the UI was also not well understood.

For the statistical assumptions questions, the weak evidence from the mid-test is surprising because the interview results were overwhelmingly positive. We surmise that this aspect may be influenced by the appropriate test selection aspect, which is the subsequent step. We revisit the latter aspect in the discussion of H2.

Using Statsplorer to prepare for learning: From the overall score in the post-test, the sequence effect is only weakly supported. The utterance counts also shows only a slight gain for Group A. These results only **weakly support H2**. Although Group A scored higher than Group B in general, the result from the reporting category was opposite: Group A scored lower than Group B in the post-test, and the Group A score had high variance. The questions in the reporting category are in remembering factual knowledge and creating procedural knowledge (Table 3). Group A also scored lower than Group B in the latter dimension. One reason is that both groups rely on Statsplorer for creating report, rather than learning to create by themselves. Group B performed better because they have been only recently exposed to Statsplorer.

For the questions concerning statistical assumptions and appropriate test selection, Group A performed better, indicating the effectiveness of Statsplorer to prepare for learning. However, the gain in the appropriate test selection category is not as clear as we expected. The reason is probably that the experimentation in the test selection was not straight-forward: In the version we used in the study, Statsplorer automatically selects most appropriate statistical test based on the results of statistical assumption tests. The users cannot change the statistical test by themselves. While this was fool-proof, it was

harder for users to experiment with the test selection mechanism. To use different tests, the users had to select different combination of variables, and switching variables changes context. As a result, the users forgot to confront their predictions about the test selection. This was reflected in the interview. Our users also asked for more freedom in selecting the tests during the interview.

Therefore, after the user study, we improved Statsplorer to allow users to select alternative statistical tests. Based on the data type and the selected variables, there are only few choices remaining. These choices are color-coded by their statistical power and by the number of many statistical assumptions the data violates. We also added a decision tree which explicitly shows the reasoning behind the test selection. It remains to be seen whether this feature will result in users trying out multiple tests to "fish" for significant results.

Limitations of the user study: There are several factors in our user study that may have caused the participants' behavior to differ from the typical usage scenario: Having to analyze the given dataset, participants may have been unfamiliar with or not motivated to exhaustively explore the data. Besides, the social dynamics of constructive interaction may have influenced their behavior, especially their focused immersion. These limitations can be addressed by observing how users analyze their own dataset, by themselves with Statsplorer.

Another limitation stems from the written tests. As shown in Table 3, the tests did not equally cover all applicable dimensions of the knowledge taxonomy. Although the tests allow us to capture the relative performance of the participants, they have not been calibrated to represent the performance of the participants at an absolute scale. Nevertheless, we provide these tests in the supplements for further replications.

LIMITATIONS AND FUTURE WORK

In order to make Statsplorer suitable for novices, we made several trade-offs in designing Statsplorer. Firstly, we focused on statistical tests that assess differences in central tendency. Other tests that are essential for HCI research, such as the chi-square test, were not implemented in Statsplorer. Limiting the set of possible statistical tests also makes it easier to determine appropriate visualizations and statistical assumption tests. In future, we plan to integrate other essential tests into Statsplorer to provide a comprehensive statistics toolkit for inexperienced HCI researchers.

Here are some possible ways to use Statsplorer as a platform for testing user interfaces for statistical analysis:

- Prepare users to graduate from Statsplorer by providing a scaffold of analysis script, e.g., in R.
- Integrate with existing tools that support other aspects of empirical HCI research, e.g., Wrangler, Touchstone.
- Evaluate alternative user interfaces for tests that are already supported in Statsplorer.
- Extend the Statsplorer back-end to support alternative statistical analysis procedures, e.g., Bayesian analysis.

CONCLUSION

We presented the design and evaluation of Statsplorer, a tool that helps inexperienced analysts both perform and learn statistical analysis. Our user study indicates that Statsplorerwas more effective in helping novices perform statistical analysis and understand the analysis process than the lecture. There was also a weak evidence that indicates having novices use Statsplorer before the lecture, allows them to prepare for future learning better than using it afterwards.

Besides empowering novices to learn statistical analysis, we hope that Statsplorer, as a platform, will inspire the development of better tools for statistical analysis.

ACKNOWLEDGMENTS

This work was funded in part by the German B-IT Foundation. We thank Pierre Dragicevic for his feedback in early iterations of Statsplorer. We thank Simon Voelker, Marty Pye, and Thorsten Karrer for their help in preparing the manuscript and the video.

REFERENCES

1. D. G. Altman. Statistics in Medical Journals: Some Recent Trends. *Statistics in Medicine*, 19(23):3275–3289, 2000.
2. American Psychological Association. *Publication Manual of the American Psychological Association*. American Psychological Association Washington DC, Sixth edition, 2006.
3. R. M. Best, M. Rowe, Y. Ozuru, and D. S. McNamara. Deep-level Comprehension of Science Texts: The Role of the Reader and the Text. *Topics in Language Disorders*, 25(1):65–83, 2005.
4. N. Boukhelifa, W. Cancino, A. Bezerianos, and E. Lutton. Evolutionary Visual Exploration: Evaluation with Expert Users. In *Proc. EuroVis '13*, 31–40. 2013.
5. J. Brandt, P. J. Guo, J. Lewenstein, M. Dontcheva, and S. R. Klemmer. Two Studies of Opportunistic Programming: Interleaving Web Foraging, Learning, and Writing Code. In *Proc. CHI '09*, 1589–1598. 2009.
6. P. Cairns. HCI... Not As It Should Be: Inferential Statistics in HCI Research. In *BCS-HCI '07*, 195–201.
7. G. Cumming. *Understanding the New Statistics: Effect sizes, Confidence Intervals, and Meta-analysis*. Routledge, 2012.
8. L. G. Daniel. Statistical Significance Testing: A Historical Overview of Misuse and Misinterpretation with Implications for the Editorial Policies of Educational Journals. *Research in the Schools*, 5(2):23–32, 1998.
9. F. D. Davis. *A Technology Acceptance Model for Empirically Testing New End-user Information Systems: Theory and Results*. Ph.D. thesis, Massachusetts Institute of Technology, 1985.

10. P. Dragicevic, F. Chevalier, and S. Huot. Running an HCI Experiment in Multiple Parallel Universes. In *Proc. CHI EA '14*, 607–618. ACM, 2014.

11. S. M. Drucker, D. Fisher, R. Sadana, J. Herron, and m. schraefel. TouchViz: A Case Study Comparing Two Interfaces for Data Analytics on Tablets. In *Proc. CHI '13*, 2301–2310. ACM, 2013.

12. R. J. Dufresne, W. J. Gerace, P. T. Hardiman, and J. P. Mestre. Constraining Novices to Perform Expertlike Problem Analyses: Effects on Schema Acquisition. *Journal of the Learning Sciences*, 2(3):307–331, 1992.

13. M. D. Dunlop and M. Baillie. Paper Rejected ($p > 0.05$): An Introduction to the Debate on Appropriateness of Null-Hypothesis Testing. *International Journal of Mobile Human Computer Interaction*, 1(3):86–93, 2009.

14. A. Fields. Discovering Statistics Using SPSS. *Beverly Hills: Sage Publications*, 2005.

15. J. Garfield. How Students Learn Statistics. *International Statistical Review/Revue Internationale de Statistique*, 25–34, 1995.

16. J. Garfield and D. Ben-Zvi. How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics. *International Statistical Review*, 75(3):372–396, 2007.

17. J. B. Garfield. Assessing statistical reasoning. *Statistics Education Research Journal*, 2(1):22–38, 2003.

18. W. D. Gray and M. C. Salzman. Damaged Merchandise? A Review of Experiments That Compare Usability Evaluation Methods. *Hum.-Comput. Interact.*, 13(3):203–261, September 1998.

19. S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proc. CHI '11*, 3363–3372.

20. M. Kaptein and J. Robertson. Rethinking Statistical Analysis Methods for CHI. In *Proc. CHI '12*, 1105–1114. 2012.

21. D. R. Krathwohl. A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, 41(4):212–218, 2002.

22. J. Lazar, J. H. Feng, and H. Hochheiser. *Research Methods in Human-Computer Interaction*. Wiley Publishing, 2010.

23. Liesbeth Kester and Paul A Kirschner and Jeroen J.G van Merrinboer and Anita Baumer. Just-in-time Information Presentation and the Acquisition of Complex Cognitive Skills. *Computers in Human Behavior*, 17(4):373–391, 2001.

24. M. C. Lovett and J. B. Greenhouse. Applying Cognitive Theory to Statistics Instruction. *The American Statistician*, 54(3):196–206, 2000.

25. W. E. Mackay, C. Appert, M. Beaudouin-Lafon, O. Chapuis, Y. Du, J.-D. Fekete, and Y. Guiard. Touchstone: Exploratory Design of Experiments. In *Proc. CHI '07*, 1425–1434. ACM, 2007.

26. J.-B. Martens. Interactive Statistics with Illmo. *ACM Trans. Interact. Intell. Syst.*, 4(1):4:1–4:28, April 2014.

27. J. E. McGrath. Human-computer Interaction. chapter Methodology Matters: Doing Research in the Behavioral and Social Sciences, 152–169. Morgan Kaufmann, 1995.

28. D. S. McNamara, E. Kintsch, N. B. Songer, and W. Kintsch. Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text. *Cognition and instruction*, 14(1):1–43, 1996.

29. J. Nielsen. *Usability Engineering*. Morgan Kaufmann, 1993.

30. J. Ooms. The OpenCPU System: Towards a Universal Interface for Scientific Computing through Separation of Concerns. *arXiv preprint*, 2014.

31. H. C. Purchase. *Experimental Human-computer Interaction: a Practical Guide with Visual Examples*. Cambridge University Press, 2012.

32. J. M. Rzeszotarski and A. Kittur. Kinetica: Naturalistic Multi-touch Data Visualization. In *Proc. CHI '14*, 897–906. 2014.

33. B. Schneider, J. Wallace, P. Blikstein, and R. Pea. Preparing for Future Learning with a Tangible User Interface: The Case of Neuroscience. *IEEE Trans. Learn. Technol.*, 6(2):117–129, April 2013.

34. D. L. Schwartz and T. Martin. Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. *Cognition and Instruction*, 22(2):129–184, 2004.

35. H. Shimazaki and S. Shinomoto. A Method for Selecting the Bin Size of a Time Histogram. *Neural Comput.*, 19(6):1503–1527, 2007.

36. J. O. Wobbrock. Practical Statistics for Human-Computer Interaction: An Independent Study Combining Statistics Theory and Tool Know-How. In *Annual Workshop of the HCI Consortium (HCIC '11)*.

37. F. W. Young and D. J. Lubinsky. Guiding Data Analysts with Visual Statistical Strategies. *J. Comp. Graph. Stat.*, 4(4):229–250, 1995.

38. A. Zieffler, J. Garfield, S. Alt, D. Dupuis, K. Holleque, B. Chang, et al. What Does Research Suggest About the Teaching and Learning of Introductory Statistics at the College Level? A Review of the Literature. *Journal of Statistics Education*, 16(2):1–23, 2008.

39. A. F. Zuur, E. N. Ieno, and C. S. Elphick. A Protocol for Data Exploration to Avoid common Statistical Problems. *Methods in Ecology and Evolution*, 1(1):3–14, 2010.