

UNVEILING THE DIGITAL AGE OF DATA SCIENCE: AN EVIDENCE-DRIVEN INVESTIGATION

By

Hannah C. Williams

Bachelor of Science – Computer Science
University of Nevada, Las Vegas
2022

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy – Computer Science

Department of Computer Science
Howard R. Hughes College of Engineering
The Graduate College

University of Nevada, Las Vegas
August 2026

Copyright by Hannah C. Williams, 2026

All Rights Reserved



Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

July 22, 2026

This dissertation prepared by

Hannah C. Williams

entitled

Unveiling the Digital Age of Data Science: An Evidence-Driven Investigation

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy – Computer Science
Department of Computer Science

Andreas Stefik, Ph.D.
Examination Committee Chair

Fatma Nasoz, Ph.D.
Examination Committee Member

Laxmi Gewali, Ph.D.
Examination Committee Member

Hal Berghel, Ph.D.
Examination Committee Member

Sarah Harris, Ph.D.
Graduate College Faculty Representative

Alyssa Crittenden, Ph.D.
*Vice Provost for Graduate Education &
Dean of the Graduate College*

Abstract

This dissertation investigates how the design of statistical software influences users' experiences, with the broader goal of advancing more accessible, user-friendly, and approachable tools for data science. While statistical software is central to education, research, and industry, many existing platforms present significant usability challenges, particularly for individuals with limited technical backgrounds, disabilities, or heightened statistical anxiety. Through a combination of literature review, empirical research, and analysis, this work examines how statistical software interface design, formats, and cognitive load interact to affect users' confidence, anxiety, comprehension and engagement. Two integral studies within the dissertation compare traditional statistical and data science software (R and Python) with experimental alternatives created using empirical design principles. Results highlight meaningful differences in user performance and emotional response across task areas such as variable selection, data programming, and output interpretation. Together, these findings inform a set of design recommendations and future directions for creating statistical tools that are not only powerful, but also intuitive, accessible, and better aligned with the needs of diverse learners and professionals in data science.

Acknowledgements

I would like to express my sincere gratitude to my dissertation committee for their valuable feedback and support throughout the development of this dissertation. I am especially thankful to my advisor, Dr. Andreas Stefik, whose guidance, mentorship, and unwavering encouragement have been instrumental at every stage of this process. I also extend my deep appreciation to my lab colleagues for their thoughtful insights, collaboration, and ongoing support.

Finally, I would like to thank my family and friends for their unwavering love, encouragement, and patience throughout this journey. Your confidence in me gave me the strength to persevere through the most challenging moments, and your support has meant more than words can express. This accomplishment would not have been possible without you.

HANNAH C. WILLIAMS

University of Nevada, Las Vegas
August 2026

Table of Contents

Abstract	iii
Acknowledgements	iv
List of Tables	x
List of Figures	xv
Chapter 1 Introduction	1
1.1 Problem	1
1.2 Research Questions	2
1.3 Impact Of Study 1	2
1.4 Impact Of Study 2	3
1.5 Significance	4
Chapter 2 Literature Review	6
2.1 Statistical Anxiety	6
2.1.1 Review Parameters	6
2.1.2 Gender and Statistical Anxiety	7
2.1.3 Education Level / Field / Experience and Statistical Anxiety	8
2.1.4 Age and Statistical Anxiety	9
2.1.5 Self-Efficacy (Confidence) and Statistical Anxiety	9
2.1.6 Performance and Statistical Anxiety	10
2.1.7 Comprehension and Statistical Anxiety	10
2.1.8 Readability and Statistical Anxiety	11
2.1.9 Disabilities and Statistical Anxiety	11
2.1.10 Attitude Towards Statistics and Statistical Anxiety	12
2.1.11 Other Socio-demographic Characteristics and Statistical Anxiety	12
2.1.12 Software and Statistical Anxiety	12

2.2 Literature Limitations	13
--------------------------------------	----

Chapter 3 Study 1: Does Learning Statistics Have to Be So Difficult? Evidence from a Randomized Controlled Trial **14**

3.1 Introduction	14
3.2 Related Work	16
3.2.1 Statistical Anxiety: Scope and Evidence	16
3.2.2 Measurement Gaps	17
3.2.3 Software-Specific Studies	17
3.2.4 Demographics and Anxiety Interaction	17
3.2.5 Direct Precursor: Terminology in Statistical Software	18
3.2.6 Summary	18
3.3 Research Question	19
3.4 Method	20
3.4.1 Inclusion and Exclusion	20
3.4.2 Participant Characteristics	20
3.4.3 Sampling Procedures	22
3.4.4 Sample Size, Power, and Precision	23
3.4.5 Measures and Covariates	24
3.4.6 Data Collection	27
3.4.7 Quality of Measurements	27
3.4.8 Instrumentation	28
3.4.9 Masking	29
3.4.10 Psychometrics	29
3.4.11 Conditions and Design	33
3.4.12 Data Diagnostics	34
3.4.13 Analytic Strategy	37
3.5 Results	39
3.5.1 Participant Flow	39
3.5.2 Recruitment	40
3.5.3 Reliability and Bias Checks	40
3.5.4 Statistics and Data Analysis	42
3.6 Qualitative Feedback	61
3.6.1 Traditional Representative Quotes	61
3.6.2 Experimental Representative Quotes	63
3.6.3 Study Format Feedback	64
3.7 Discussion	65

3.7.1	Primary Findings: Software Design Effects	65
3.7.2	Secondary Analyses and Broader Patterns	65
3.7.3	Uncertainty as a Distinct Construct	67
3.7.4	Interpretation: Where and Why Design Helped	67
3.7.5	Generalizability	68
3.7.6	Implications	70

Chapter 4 Study 2: Learning to Swim in a Sea of Statistics Anxiety: A Randomized Controlled Trial

		72
4.1	Introduction	72
4.1.1	What is Declarative Data Science?	73
4.2	Hypotheses, Aims, and Objectives	74
4.3	Method	75
4.3.1	Inclusion and Exclusion	75
4.3.2	Participant Characteristics	75
4.3.3	Sampling Procedures	75
4.3.4	Sample Size, Power, and Precision	78
4.3.5	Measures and Covariates	78
4.3.6	Data Collection	80
4.3.7	Quality of Measurements	80
4.3.8	Instrumentation	80
4.3.9	Masking	90
4.3.10	Psychometrics	91
4.3.11	Conditions and Design	92
4.3.12	Data Diagnostics	94
4.3.13	Analytic Strategy	94
4.4	Results	95
4.4.1	Engagement	95
4.4.2	Performance	97
4.4.3	Affective Outcomes	101
4.4.4	Exploratory Analyses	105
4.5	Qualitative Feedback	106
4.5.1	Task-Level Feedback	107
4.5.2	End-of-Study Feedback	108
4.6	Discussion	109
4.6.1	Support of Original Hypotheses	109
4.6.2	Interpretation	110

4.6.3	Generalizability	112
4.6.4	Implications	113
Chapter 5	Declarative Layer for Statistical and Machine Learning Workflows	116
5.1	Introduction	116
5.2	Declare a Distinct Identifier	117
5.3	Declare Input Variable(s)	118
5.4	Declare Input Interaction(s)	118
5.5	Declare Output Variable(s)	119
5.6	Declare an Analytical Goal	120
Chapter 6	Implementation and Mathematical Foundations	121
6.1	Descriptive Statistics	122
6.1.1	Measures of Central Tendency	123
6.1.2	Measures of Variability	125
6.1.3	Measures of Shape	126
6.1.4	Measures of Association	128
6.1.5	Future Additions	132
6.2	Inferential Statistics	133
6.2.1	Tests for Associations	134
6.2.2	Tests For Differences	139
6.2.3	Assumption Tests	166
6.2.4	Routing Logic	166
6.3	Predictive Modeling	170
6.3.1	Supervised Learning	170
6.3.2	Routing Logic	171
6.4	Scope of This Chapter	172
Chapter 7	Conclusion	173
7.1	Downstream Opportunities	174
7.2	Closing	176
Appendix A	Study 1	177
A.1	Tasks	177
A.2	Statistical Analysis Result Tables	190
A.2.1	Measure Validation	190
A.2.2	Study-Wide	191
A.2.3	Selection Area	193

A.2.4	Input Area	195
A.2.5	Output Area	198
Appendix B	Study 2	204
B.1	Tasks	204
Bibliography		222
Curriculum Vitae		236

List of Tables

Table 3.1	Participant distribution	21
Table 3.2	Overview of the selection tasks. See Appendix A.1 for full details.	28
Table 3.3	Overview of the input tasks. See Appendix A.1 for full details.	29
Table 3.4	Overview of the output tasks. See Appendix A.1 for full details.	31
Table 3.5	Grading criteria for evaluating task 3	34
Table 3.6	Grading criteria for evaluating task 4	35
Table 3.7	Participant characteristics re-categorized	39
Table 3.8	ANCOVA results for study-wide average anxiety.	44
Table 3.9	ANCOVA results for study-wide average confidence.	45
Table 3.10	ANCOVA results for study-wide average uncertainty.	46
Table 3.11	ANCOVA results for study-wide average correctness after removing unsure responses.	47
Table 3.12	ANOVA results for perceived task difficulty.	48
Table 3.13	Post-hoc comparisons for design and task area on perceived task difficulty	48
Table 4.1	Participant demographic characteristics (N = 120)	76
Table 4.2	Statistical methods and key concepts across the 12 tasks.	81
Table 4.3	Experimental interface block palette information tooltip provided.	86
Table 4.4	Control interface block palette variable block information tooltip provided.	87
Table 4.5	Control interface block palette pingouin test/model block information tooltip provided.	88
Table 4.6	Control interface block palette statsmodels test/model block information tooltip provided.	89
Table 4.7	Control interface block palette scikitlearn test/model block information tooltip provided.	89
Table 4.8	Priority-ordered feedback message system. Upon each submission, the participant's answer was evaluated against the hierarchy below and only the highest-priority applicable message displayed.	91
Table 4.9	Number of scored criteria by task and group	92
Table 4.10	Scoring criteria by component and group	93
Table 4.11	Mixed ANOVA for effort-weighted engagement	95

Table 4.12	Post-hoc comparisons for effort-weighted engagement by task (Control – Experimental)	96
Table 4.13	Mixed ANOVAs for first-attempt and first-intent correctness	98
Table 4.14	Post-hoc comparisons for first-attempt correctness by task (Control – Experimental) .	100
Table 4.15	Post-hoc comparisons for first-intent correctness by task (Control – Experimental) . .	101
Table 4.16	Mixed ANOVA for anxiety ratings	103
Table 4.17	Mean anxiety and confidence by group and task outcome	103
Table 4.18	Mixed ANOVA for confidence ratings	105
Table 4.19	Post-hoc comparisons for confidence by task (Control – Experimental)	105
Table 6.1	Contingency table for the chi-squared test of independence.	138
Table 6.2	Contingency table for the McNemar and McNemar-Bowker test of symmetry.	141
Table 6.3	Binary-encoded outcome table for Cochran’s Q test of differences.	142
Table 6.4	Example encoding of a three-level categorical input using the ReferenceEncoder.	158
Table 6.5	Example encoding of an interaction between a three-level and a two-level categorical input.	158
Table 6.6	GetTest routing: no output declared	167
Table 6.7	GetTest routing: one numerical output	167
Table 6.8	GetTest routing: multiple numerical outputs	168
Table 6.9	GetTest routing: one ordinal output	168
Table 6.10	GetTest routing: one categorical output	168
Table 6.11	GetTest routing: no input, multiple outputs (pairwise correlation matrix)	169
Table A.1	Independent grading agreement between Rater A and Rater B for a subset of code responses. The values shown were used to calculate Cohen’s kappa reported in the main text, providing a measure of inter-rater reliability.	190
Table A.2	Test-retest correlations between participants’ first and second task ratings in each area. Higher correlations indicate greater consistency of participants’ responses across similar tasks, providing evidence of the reliability of these measures. Overall, the results suggest reasonable reliability for both anxiety and confidence ratings in all areas.	191
Table A.3	Correlation between participants’ average anxiety and confidence ratings across the study, task areas and aspects. Consistent with prior research, anxiety and confidence were moderately negatively correlated, supporting the validity of the measurement system.	192
Table A.4	ANOVA results for anxiety and confidence on neutral data- and information-related items. No significant main effects of design were observed. The design × gender interaction showed females in the Traditional group were slightly more anxious and less confident than males; controlling for academic field reduced this effect.	193

Table A.5	ANOVA results for performance on neutral data- and format-related questions. No significant main effects of design were observed. The design $\times$ gender interaction showed females in the Traditional group were slightly more uncertain than males; controlling for academic field reduced this effect.	193
Table A.6	Post hoc Tukey–Kramer tests for study-wide analyses. Significant pairwise differences follow up the ANCOVA results reported in the main text. This indicates that participants who took at least one statistics course generally showed higher correctness and lower uncertainty and those who took multiple felt more confident as well. Gender differences in confidence were often related to academic field, participants in the formal sciences were most confident, and females in other fields were significantly less confident than all other participants. Post hoc comparisons are shown only for significant effects that involve more than two groups; effects with only two groups are reported in the main text.	194
Table A.7	ANCOVA results for selection area uncertainty (proportion of “not sure” responses across all questions) and correctness (calculated after excluding “not sure” responses). The design reduced uncertainty, indicating participants were more decisive when selecting statistical tests. Statistics experience also predicted lower uncertainty, yet not higher correctness, suggesting increased decisiveness did not equate to greater accuracy. Gender influenced uncertainty (men were more uncertain) and academic field influenced accuracy (formal sciences were more correct).	195
Table A.8	Selection area ANCOVA results for overall and nomenclature-related ratings of anxiety and confidence. Design did not significantly affect participants’ overall anxiety or confidence, but it did yield a positive effect on confidence when ratings were limited to nomenclature-related items (i.e., subs-scales reflecting participants’ reactions to statistical test names during task selection). This suggests that while overall perceptions were similar across groups, the experimental design improved participants’ confidence specifically in navigating statistical terminology. Several effects emerged for covariates (e.g., disability status, statistics course experience, age).	196
Table A.9	Tukey–Kramer tests for selection area. The gender gap in confidence did not persist when considering only nomenclature-related items, suggesting more of a data-related difference. Having taken more than one statistics course consistently helped promote higher confidence in the selection area, and having taken at least one course helped promote engagement and decisiveness.	197

Table A.10	ANCOVA results for input area correctness (proportion correct after removing “not sure” responses). The experimental design improved correctness, with participants in the experimental group performing more accurately than those in the traditional group. Academic field also predicted higher correctness, whereas other covariates (e.g., gender, age, disability, statistics course experience) had minimal impact. Uncertainty was measured in the input area, but because only two coding questions were used, values were limited to 0%, 50%, or 100% and were not suitable for ANOVA analysis; see the main text for the descriptive results of input area uncertainty.	198
Table A.11	Input area ANCOVA results for overall and code-related ratings of anxiety and confidence. Design had a positive effect on participants’ confidence across both overall and code-related ratings and reduced anxiety specifically for code-related items (e.g., format, naming convention, amount and syntax of the code), indicating that the experimental design helped participants feel more confident and less anxious when entering statistical code. Several effects emerged for covariates: gender and age were associated with overall anxiety, while statistics course experience and academic field were linked to code-related confidence and anxiety. These results suggest that design improvements can selectively enhance participants’ experience with code entry tasks, particularly in reducing anxiety related to understanding and using statistical code.	199
Table A.12	Tukey–Kramer post hoc comparisons for input area. Having taken at least one statistics course consistently helped promote higher confidence in the input area. Women outside of formal sciences felt significantly less confident than all other participants.	200
Table A.13	Task 3 correctness when attempted for each criterion. See the Psychometrics section for a full description.	200
Table A.14	Task 4 correctness when attempted for each criterion. See the Psychometrics section for a full description.	201
Table A.15	Output area ANCOVA results for overall ratings of anxiety and confidence. Design significantly affected anxiety and confidence, with experimental participants showing lower anxiety and higher confidence. Covariate effects were mostly nonsignificant. . . .	202
Table A.16	ANCOVA results for output area uncertainty (proportion of “not sure” responses across all questions) and correctness (calculated after excluding “not sure” responses). The experimental design reduced uncertainty and improved accuracy. Confidence strongly predicted both, while course experience reduced uncertainty but was less clearly tied to accuracy.	203
Table B.1	Task 1 dataset excerpt (first 8 rows of 30).	204
Table B.2	Task 2 dataset excerpt (first 9 rows of 30).	206
Table B.3	Task 3 dataset excerpt (first 9 rows of 30).	207

Table B.4	Task 4 dataset excerpt (first 8 rows of 20).	208
Table B.5	Task 5 dataset excerpt (first 9 rows of 30).	210
Table B.6	Task 6 dataset excerpt (first 9 rows of 30).	211
Table B.7	Task 7 dataset excerpt (first 8 rows of 40).	212
Table B.8	Task 8 dataset excerpt (first 8 rows of 40).	214
Table B.9	Task 9 dataset excerpt (first 8 rows of 30).	215
Table B.10	Task 10 dataset excerpt (first 6 rows of 90).	217
Table B.11	Task 11 dataset excerpt (first 6 rows of 40).	218
Table B.12	Task 12 dataset excerpt (first 6 rows of 36).	220

List of Figures

Figure 3.1	Experimental task 2 - question 6	28
Figure 3.2	Traditional task 2 - question 6	29
Figure 3.3	Experimental task 3 - example code	30
Figure 3.4	Traditional task 3 - example code	30
Figure 3.5	Experimental task 5 - output	32
Figure 3.6	Traditional task 5 - output	33
Figure 3.7	Experimental task 3 criteria. Numbers indicate which evaluation criterion (shown in Table 3.5) was applied to that portion of the code.	36
Figure 3.8	Traditional task 3 criteria. Numbers indicate which evaluation criterion (shown in Table 3.5) was applied to that portion of the code.	37
Figure 3.9	Experimental task 4 criteria. Numbers indicate which evaluation criterion (shown in Table 3.6) was applied to that portion of the code.	38
Figure 3.10	Traditional task 4 criteria. Numbers indicate which evaluation criterion (shown in Table 3.6) was applied to that portion of the code.	38
Figure 3.11	Correctness (green) and uncertainty (blue) rates on neutral data-related questions. Performance was similar between Experimental and Traditional groups, indicating that software design did not bias responses on questions unrelated to design.	41
Figure 3.12	Completion rate at each stage of the study	43
Figure 3.13	Perceived difficulty for each task area. Participants rated how challenging the two tasks in each area were on a 5-point scale (1 = very easy, 5 = very difficult). The figure highlights that participants using the Experimental design rated input and output tasks as less difficult than those using the Traditional design, while selection tasks showed no difference.	57

Figure 3.14	Correctness (green) and uncertainty (blue) rates on design-dependent questions. Uncertainty was much higher in the Traditional group and correctness was similar between Experimental and Traditional groups in two of the three questions, indicating that the naming convention can promote engagement but needs more thoughtful and specific consideration to also improve accuracy.	58
Figure 3.15	Correctness (green) and uncertainty (blue) rates for coding tasks. Overall, the Experimental group showed slightly higher correctness and slightly lower uncertainty across both tasks compared to the Traditional group. This pattern suggests that the Experimental coding syntax supports engagement and improves first-pass accuracy when participants attempt the tasks.	59
Figure 3.16	Correctness (green) and uncertainty (blue) rates for output interpretation questions. Overall, the Experimental group showed higher correctness across most concepts, while the Traditional group exhibited higher uncertainty. One exception was interpreting t-test significance from a p-value, where correctness was similar across groups. These results suggest that the Experimental output design improves engagement and supports more accurate interpretation of statistical results.	60
Figure 4.1	The full task interface, with task description (A), dataset (B), block palette (C), and workspace (D) labeled. The text content in C and D differ between the Control (shown here) and Experimental interfaces.	82
Figure 4.2	The experimental interface during Task 1, showing feedback after an incorrect submission. The submitted variable block is highlighted green (correct) and the submitted test block red (incorrect). An additional unsubmitted variable block (gray) is visible in the workspace, and the second dropdown is open showing available role options. . .	84
Figure 4.3	The control interface during Task 1, showing feedback after an incorrect submission. The variable block remains gray (<i>astype</i> variable blocks do not change color in the control interface) and the test block is highlighted red indicating an incorrect submission. The information tooltip for the chi-squared test block is open, displaying a brief description of the test.	85
Figure 4.4	The per-task ratings screen shown after task completion, with both anxiety and confidence ratings selected and the ratings panel re-opened for review. Anxiety is rated on a seven-point scale from “No anxiety at all” to “An extreme amount of anxiety”; confidence uses an identical scale. An open-ended feedback field is also provided. . . .	90
Figure 4.5	Effort-weighted engagement by task and group. The upper portion of the chart reflects successful task completion, the lower portion reflects task abandonment. Violin plots show the full distribution; points and error bars represent group means with 95% confidence intervals.	97

Figure 4.6	First-attempt and first-intent correctness by task (give-ups scored as incorrect).	99
Figure 4.7	Correctness by task, conditional on attempting the task.	102
Figure 6.1	Experimental interface pseudocode for generating a statistical summary.	122
Figure 6.2	Experimental interface pseudocode for generating an inferential test.	133
Figure 6.3	Experimental interface pseudocode for generating an inferential test.	170
Figure A.1	Task 1 (Experimental version): Selection. Correct answers – weight and tree height: quantitative; color and taste: qualitative; first row same apple: yes; test for weight vs. tree height: <i>CorrelateGroups()</i> ; test for color vs. taste: <i>CompareCounts()</i>	178
Figure A.2	Task 1 (Traditional version): Selection. Correct answers – weight and tree height: quantitative; color and taste: qualitative; first row same apple: yes; test for weight vs. tree height: <i>cor.test()</i> ; test for color vs. taste: Traditional = <i>chisq.test()</i>	179
Figure A.3	Task 2 (Experimental version): Selection. Correct answers – species: qualitative; bpm: quantitative; species as predictor, bpm as outcome; repeated measurements: no; test for comparing bpm across species: <i>CompareMeans()</i>	180
Figure A.4	Task 2 (Traditional version): Selection. Correct answers – species: qualitative; bpm: quantitative; species as predictor, bpm as outcome; repeated measurements: no; test for comparing bpm across species: <i>aov()</i>	181
Figure A.5	Task 3 (Experimental version): Input. Correct answers – file name: <i>"heartrates.csv"</i> ; input factor: <i>"species"</i> ; output variable: <i>"bpm"</i> ; main test: <i>CompareMeans()</i>	182
Figure A.6	Task 3 (Traditional version): Input. Correct answers – file name: <i>"heartrates.csv"</i> ; input factor: <i>as.factor(frame\$species)</i> ; output variable: <i>bpm</i> ; main test: <i>aov(bpm~species, data=frame)</i>	183
Figure A.7	Task 4 (Experimental version): Input. Correct answers – file name: <i>"medication.csv"</i> ; repeated factor variable: any name in first parameter of <i>frame:SetRepeatedFactor("...", "...")</i> ; repeated factor levels: <i>"after,during,before"</i> in the second parameter of <i>frame:SetRepeatedFactor("...", "...")</i> ; output variable: generated automatically from <i>frame:SetRepeatedFactor("...", "...")</i> ; main test: <i>CompareMeans()</i> . See Psychometrics (Section 3.4.10) for full breakdown of grading criteria.	184
Figure A.8	Task 4 (Traditional version): Input. Correct answers – file name: <i>"medication.csv"</i> ; repeated factor variable: any consistent name set in the second parameter (<i>key</i>) of <i>gather(...)</i> ; repeated factor levels: <i>"after","during","before"</i> in the fourth, fifth and sixth parameters of <i>gather(...)</i> ; output variable: any consistent name set in the third parameter (<i>value</i>) of <i>gather(...)</i> ; subject Identifier: <i>as.factor(long_frame\$person)</i> ; main test: should use all consistent variable and object names in <i>aov(...)</i> . See Psychometrics (Section 3.4.10) for full breakdown of grading criteria.	185

Figure A.9 Task 5 (Experimental version): Output. Correct answers – normality assumption: yes; sample sizes equal: no; statistical significance: no; Cohen’s d: small.	186
Figure A.10 Task 5 (Traditional version): Output. Correct answers – normality assumption: yes; sample sizes equal: no; statistical significance: no; Cohen’s d: small.	187
Figure A.11 Task 6 (Experimental version): Output. Correct answers – homogeneity of variance: yes; normality: yes; group effect significant: yes; effect size (eta-squared): medium; Tukey post hoc: group1 vs. group3.	188
Figure A.12 Task 6 (Traditional version): Output. Correct answers – homogeneity of variance: yes; normality: yes; group effect significant: yes; effect size (eta-squared): medium; Tukey post hoc: group1 vs. group3.	189

Chapter 1

Introduction

1.1 Problem

In our world, which is increasingly driven by data, the ability to correctly analyze, interpret, and apply statistical methods has become ever-more critical [16]. From scientific research and medical or pharmaceutical decision-making to policy development and corporate analytics, statistical reasoning is necessary for the most vital aspects of modern decision making. In critical settings, such as healthcare, finance, and engineering, errors in statistical analysis can lead to dire consequences, from misinformed policy decisions to flawed scientific conclusions. However, even outside of these critical fields, statistical comprehension still remains a significant challenge for researchers, politicians, and the general public alike [2, 83]. Political discourse, for example, frequently relies on statistical data, yet misinterpretations or manipulations of statistical findings can shape public opinion in misleading ways.

Despite needing statistical understanding, mastering statistical concepts and analyses remains a significant barrier for many. Research has consistently shown that statistics is one of the most difficult subjects for students across all levels of education [101]. Even experienced researchers often struggle with the complexity of statistical methods, particularly when it comes to selecting appropriate tests, interpreting results, and ensuring their findings are replicable. As our reliance on data science and statistical software grows, so does the challenge of making these tools accessible, interpretable, and free from errors that could compromise research integrity.

Statistical software has revolutionized how researchers conduct data analysis, but it has also introduced new challenges. As researchers become more dependent on software for statistical computations, programming errors, software limitations, and usability issues become increasingly consequential. A misinterpretation of statistical output due to confusing software design can lead to flawed conclusions, and research results could be affected by software bugs. For instance, during the development of the experimental statistical software for this dissertation, two critical bugs were discovered in well-known statistical software (R and

Apache Commons) that led to incorrect p-values in some comparisons. These errors highlight a broader concern: if experienced researchers can encounter such issues, how many unnoticed errors exist in widely used statistical software and what sort of research has been done on the topic?

Furthermore, steep learning curves, confusing syntax or poorly structured output can force researchers to rely on professional statisticians, pay large fees for statistical consulting and expensive software licenses, or limit the variety of their research designs to only those they feel comfortable analyzing. This reliance not only creates economic and accessibility barriers but also restricts diverse scientific experiments and data analysis.

With statistical analysis comes statistical anxiety, which has been widely studied and numerous scales have been developed to measure it [29, 56, 139, 18]. However, existing research on statistical anxiety overwhelmingly focuses on academic settings, particularly university students learning the mathematical operations. Far less attention has been given to statistical anxiety among professional researchers, high school students, or individuals working in industries where data analysis is becoming an essential skill.

Furthermore, validated prominent statistical anxiety scales do not account for the role of software in statistical anxiety. Given that modern statistical analysis is almost entirely dependent upon software, it is surprising that little research has been done to investigate how the structure, syntax, and output of statistical software contribute to user anxiety. This dissertation aims to address this gap by considering how software usability and design influence not only statistical performance but also anxiety, confidence, and user comprehension.

1.2 Research Questions

1. How does statistical software design impact users' ability to perform statistical analyses accurately and confidently?
2. How do software design changes influence anxiety levels and to what extent does statistical software contribute to statistics anxiety?
3. How can statistical software be improved to enhance user experience while maintaining accuracy and replicability?

1.3 Impact Of Study 1

The first study in this dissertation (Chapter 3) serves as an initial exploration of the complex interactions between statistical software design, user experience, and statistical anxiety. This study compares an experimental statistical software approach (Experimental) to a standard approach (Traditional) consisting of popular packages from the widely used open-source software R, to identify areas where changes to software design can improve usability and reduce anxiety while maintaining essential research principles such as

replicability.

Some key findings from this study include:

- Statistical output interpretation is the most difficult and anxiety-inducing area for users. Simple design changes can significantly enhance comprehension and reduce perceived difficulty. The changes included formatting output in sentence structures using around an 8th-grade reading level [71] and accompanying numerical results in APA format and in tables.
- Renaming statistical tests in a more intuitive manner improves user understanding of which test to use and promotes user engagement. However, this requires careful consideration, as names must be both intuitive and accurately descriptive of the statistical method rather than just a description of mathematical functions. This study tested the effects of the experimental names proposed by Kluthe et. al [72] for three statistical tests to gauge potential improvements in usability.
- Programming syntax affects statistical anxiety and programming performance. This study specifically employed code for ANOVA and repeated measures ANOVA, revealing that simple improvements and consistency in software syntax can assist building users' self-efficacy, reduce anxiety, increase engagement, and improve performance in statistical programming tasks.

1.4 Impact Of Study 2

The second study in this dissertation (Chapter 4) builds directly on the findings of Study 1 by shifting from surface-level interface simplifications to structural automation. Where Study 1 demonstrated that modest changes to test naming, syntax, and output format could improve confidence and engagement, its results also revealed that surface simplification alone is insufficient: analytical accuracy showed mixed results and anxiety persisted, suggesting that many difficulties in applied statistics stem not from conceptual misunderstanding but from the procedural complexity imposed by the tools themselves. Study 2 introduces *Declarative Data Science*, a semantic, structure-first approach in which users explicitly declare variable types, analytical roles, and structural relationships, after which test selection, encoding, and preprocessing follow deterministically. This relocates cognitive effort away from procedural decisions and toward articulating the structure of the research design itself.

Some key findings from this study include:

- Requiring explicit semantic declaration of variable types, roles, and structural relationships substantially improved engagement, accuracy, and confidence relative to a traditional workflow. The experimental interface carried an initial learning curve: control participants held an advantage on the earliest tasks, where familiarity with conventional workflows and terminology provided a head start. However, experimental participants adapted quickly, and as task complexity increased their engagement, accuracy, and confidence all grew and surpassed the control group. Control participants, by contrast,

relied on recalled statistical knowledge and procedural familiarity, strategies that proved unsustainable as task complexity increased and existing knowledge ran out, leading to disengagement and sinking confidence.

- Anxiety differences between groups were not substantial at a glance, but examining anxiety in the context of task outcome revealed two meaningfully distinct low-anxiety states. For experimental participants, low anxiety reflected genuine resolution: it accompanied success and moved in the expected direction as participants progressed toward mastery. For a substantial portion of control participants, however, low anxiety reflected something different; an affective collapse following complete disengagement. These participants reported anxiety levels indistinguishable from those who succeeded, yet their confidence remained near the floor. This pattern we refer to as an *anxiety-to-apathy* transition, in which unresolvable difficulty leads to disengagement rather than ongoing distress [124], and it was substantially more prevalent in the control group.

1.5 Significance

This dissertation aims to advance both practical and theoretical knowledge in statistics education, data science, and software usability by addressing key challenges in statistical software design and user experience. As statistical literacy becomes increasingly essential across disciplines, ensuring that statistical tools are intuitive and accessible can improve research accuracy, data-driven decision-making, and educational outcomes.

By identifying and addressing pain points in statistical software design, this research contributes to the development of more effective, user-friendly tools that minimize barriers to statistical analysis. Statistical software should not only facilitate accurate computations but also enhance users' understanding of data, reducing trial-and-error attempts or reliance on external expert consultation. This paper highlights ways to mitigate statistical anxiety by improving software structure, simplifying output interpretation, and refining syntax to be more intuitive for users at all levels.

Beyond individual software improvements, this research has broader implications for the restructuring of inferential and descriptive statistics, as well as the standardization of statistical reporting and visualizations. A more consistent and accessible statistical experience can improve data transparency, replicability, and interdisciplinary collaboration, ensuring that statistical analysis remains a tool for insight rather than an obstacle to overcome.

Additionally, these findings can inform educators and researchers on how to better support individuals struggling with statistical concepts. By aligning software design with best practices in education and cognitive accessibility, this study can help shape future approaches to teaching statistics, in academic settings and in professional environments, where data analysis is crucial.

Ultimately, as we become more dependent on statistical analysis in a wide range of fields, it is crucial

to ensure that the tools we use enhance rather than hinder our ability to conduct sound, reliable, and comprehensible research.

Chapter 2

Literature Review

2.1 Statistical Anxiety

Statistics anxiety has shown to be a significant barrier for learners at nearly all educational levels, from high school to graduate studies, even onto professionals [83, 68], impacting their attitudes, performance, confidence, and comprehension. This anxiety is particularly pronounced among students in non-mathematical fields. Research indicates that approximately 75% to 80% of graduate students experience uncomfortable levels of statistical anxiety. Unfortunately, this phenomenon extends beyond graduate programs; many undergraduate students are required to take statistics and quantitative research courses, with up to 80% reporting high levels of anxiety while enrolled [101]. A substantial number of students identify statistics courses as the most anxiety-inducing component of their academic programs [86].

The impact of instructional format on statistical anxiety has also been explored. Frey-Clark and Nateson studied about 300 freshman undergraduates comparing online and traditional in-person instruction, and found that while online formats initially resulted in higher reported anxiety, both formats produced similar anxiety levels by the end of the course [49]. Furthermore, elevated levels of statistical anxiety have been linked to academic dishonesty, particularly in online learning environments [40, 39]. Literature reviews and developed scales have identified six key dimensions of statistical anxiety: Interpretation Anxiety, Test and Class Anxiety, Fear of Asking for Help, Worth of Statistics, Computation Self-Concept, and Fear of Statistics Teachers. The first three measures focus on anxiety regarding statistical tasks, while the latter three reflect attitudes and perceptions toward statistics [61, 53, 86].

2.1.1 Review Parameters

Several detailed literature reviews on statistical anxiety have been conducted over the years, and this paper draws on key reviews from Onwuegbuzie and Wilson (2003) [101], Baloglu and Zelhart (2003) [5], Chew and Dillon (2014) [18], Macher et al. (2015) [85], Ralston et al. (2016) [113], Ghani and Maat (2018) [53], and the

latest done by Trassi et al. (2022) [134] which surveyed about 200 papers regarding the topic. In addition to these extensive works, a search for relevant literature was conducted using the extensive databases available through the university library. The search was limited to English, online, full-text, peer-reviewed articles that included key terms such as “statistical anxiety” or “statistics anxiety,” which initially yielded over 250 results. Given that the most recent comprehensive literature review on the topic was completed in 2022 [134], an additional filter was applied to include only articles published after 2022 with available abstracts. This refinement narrowed the selection to 61 articles. From these, only those providing evidence and directly relevant to the current research were included, leaving approximately 40 articles on statistical anxiety for this review, in addition to the previous systematic reviews.

The studies reviewed span a wide range of countries and academic disciplines, though most of them involved students from social sciences and health sciences fields. The participants in these studies ranged from high school to post-graduate levels, including undergraduate, graduate, doctoral, and post-doctoral students. To further explore how different statistical software impacts statistical anxiety, additional searches were conducted using keywords like “R,” “SPSS,” “Excel,” and other software alongside “anxiety” and “statistics.” These searches, limited to the last 10 years, revealed varying results: “R” produced four relevant articles, “SPSS” yielded three, “Excel” produced one, and searches for “Python,” “MiniTab,” and “MATLAB” returned no relevant studies. This range of findings provides a diverse foundation for exploring the impact of software on statistical anxiety across different educational and professional contexts.

2.1.2 Gender and Statistical Anxiety

Research consistently indicates that women tend to report higher levels of statistical anxiety than men across various contexts, including coursework, exam preparation, and overall attitudes toward statistics [11, 47, 37, 20, 90]. Conversely, men have been found to experience greater anxiety when seeking help with statistics [87, 84]. Age may further influence these patterns, as older men report higher levels of statistical anxiety than younger men, a trend not observed in women [113], a topic explored further in 2.1.4.

While gender differences in statistical anxiety are widely reported, findings are not entirely consistent. For example, Mji (2009) [95] found no significant gender gap between 226 technology students (mostly women), and Kian et al. (2022) reported that postgraduate women experienced lower anxiety levels than their male peers [68] in a relatively small qualitative study involving 36 postgraduate students. These inconsistencies might suggest that gender alone does not determine statistical anxiety; instead, factors such as education level, academic field, and personal experience with statistics or math play significant roles [3, 37, 90]. Despite frequently reporting higher anxiety, women often perform just as well as men on statistical tasks [87] or better [84], which is discussed further in 2.1.6.

The systematic review by Trassi et al. found that the majority of statistical anxiety studies have been conducted with predominantly female participants, largely because over 80% of the samples were drawn from human science and health fields, where women are more prevalent. In contrast, only 15% of studies included

participants from STEM fields [134]. This suggests that academic field may also contribute to differences in statistical anxiety, a relationship explored further in 2.1.3. Additionally, Ralston’s review (2016) found that some research suggests that men may down-play their feelings of anxiety, leading to lower self-reported scores and potentially exaggerating the apparent gender gap [112].

Women have been found to experience higher levels of programming and computer anxiety, which may further exacerbate statistical anxiety, particularly when statistical software is involved [114]. However, further research is needed to determine the extent to which technological anxiety contributes to overall statistical anxiety, a topic examined in 2.1.12. While gender appears to influence statistical anxiety to some extent, its effects are intertwined with other factors such as education level, academic field, age, statistical performance, and familiarity with technology. Although findings are sometimes conflicting, the majority of studies suggest that women report higher statistical anxiety than men [134].

Gender’s association with differences in other perceptions related to statistics, like attitude towards statistics [37], and other socio-demographic characteristics, like age [37], academic background or level [90], shows that the association between gender and statistical anxiety is more complex than early research suggested.

2.1.3 Education Level / Field / Experience and Statistical Anxiety

Education level generally correlates with statistical anxiety, with graduate students and professionals typically experiencing lower levels of math and statistics anxiety than undergraduate students [90]. However, even experienced users report varying degrees of statistical anxiety, often influenced by their familiarity with concepts, terminology, and software [115, 47, 142], though not all studies have shown a difference and academic field may again play a role. Birenbaum (Israel, 1994) found that experience with statistics courses did not affect statistics anxiety in female undergraduate education science students, but failed to include males in the study to determine if the effect is the same [12], and Beurze (Netherlands, 2014) found statistics course experience had no effect on anxiety levels among 520 medical students (67% female and 75% having never taken a stats course). [11]. Suggesting that factors beyond mere exposure, such as instructional methods and software usability, may play a critical role in shaping users’ confidence and comfort with statistical analysis.

Research has consistently found that individuals in mathematics and formal sciences exhibit lower statistical anxiety compared to those in other fields, this suggests that propensity to mathematics or related domains in conjunction with statistics experience may have a positive effect in limiting anxiety [93, 95, 134]. In other words, individuals with lower math skills or weaker academic performance in math tend to report the highest levels of anxiety and most negative behaviors [119, 40], whereas those with stronger mathematical backgrounds generally have a more positive attitude toward statistics [92], perform better, and experience less anxiety. A relationship between experience and age has also been observed and is discussed further in the next section 2.1.4.

2.1.4 Age and Statistical Anxiety

Earlier research like Mustafa Baloglu’s 2003 study of the statistical anxiety across a variety of college students ($n = 246$, predominately females in social sciences) conducted to gauge its relationship with various socio-demographic characteristics [3], found that when accounting for experience as a covariate, older students had the most positive attitudes toward statistics, but showed significantly higher statistics anxiety than younger students. Specifically, older women had the most statistics anxiety, whereas older men experienced had the least statistics anxiety. However later studies and reviews have come to different results. Ralston’s 2016 literature review indicated that while young women tend to experience greater statistical anxiety than young men, this pattern shifts with age, older women reported lower anxiety levels than older men. Additionally, older men exhibited higher statistical anxiety than younger men, whereas no significant differences were observed between younger and older women [113].

Furthermore, not all research supports a clear relationship between age and statistical anxiety. Some studies found no significant differences in anxiety levels based on age [11], and others (Edirisooriya, 2021) concluded that too little research as been done on older participants, predominately those under 30 years old, to diagnose this relationship [37], suggesting that the relationship between age, gender, and statistical anxiety is highly influenced by multiple factors.

These ever-conflicting results show that the associations between age and statistics anxiety is quite complicated and nuanced. Additionally, a review by Trassi et al. (2022) highlighted a sizable gap in research investigating the effects of age and other socio-demographic factors on statistical anxiety, emphasizing the need for further research in this area [134].

2.1.5 Self-Efficacy (Confidence) and Statistical Anxiety

Finney and Schraw (2003) [44] defined statistical self-efficacy two ways, current self-efficacy: “confidence in one’s abilities to solve specific tasks related to statistics” and self-efficacy to learn: “confidence in one’s abilities to learn the skills necessary to solve specific tasks related to statistics.” For the context of this dissertation, the first definition is highlighted. The scales they created to measure self-efficacy use the term ‘confidence’ throughout and thus the present work may use the terms interchangeably at times. In other words, the term confidence in this work is referring to self-efficacy in statistics related tasks. Research has consistently demonstrated a significant correlation between statistical anxiety and statistical self-efficacy [12, 108, 90, 47, 134]. This correlation suggests that individuals who experience greater levels of anxiety regarding statistics tend to have lower confidence in their ability to understand and apply statistical concepts effectively. Conversely, those with higher levels of statistical self-efficacy tend to experience less anxiety when engaging with statistical tasks, as they feel more capable and confident in their abilities to handle statistical challenges [108, 134]. Confidence (self-efficacy) is also shown to influence effort and performance during statistical learning [87, 66].

2.1.6 Performance and Statistical Anxiety

In the research into this relationship, performance is often defined and calculated using course grades, GPA, or test scores. Contrary to statistical anxiety, trait anxiety (i.e., general proneness to anxiety) may actually help motivate students to apply more effort into learning [63, 53, 87, 85, 84]. While an optimal level of trait anxiety may have a positive effect on a student's performance, individuals experiencing high levels of statistical anxiety may exhibit impaired cognitive functioning, reduced problem-solving abilities, and decreased attentional focus, all of which can hinder their performance on statistical tasks. Moreover, heightened statistics anxiety may lead to avoidance behaviors or procrastination, further exacerbating performance issues [87, 103, 134]. This corresponds with the result of a similar studies conducted up to a decade earlier [101, 53, 11, 152, 115]. However, there is some indication that self-efficacy may be a better predictor of performance than anxiety [59] as some studies did not find relationships between statistical anxiety and performance [134]. Additionally, other negative behaviors, like academic dishonesty, observed by Sandoz (2017) in 132 social and human science students and Eshet (2021,2022) in 291 predominately female social science students, appeared to help students mediate statistical anxiety and regulate the negative impact it would have on performance [119, 5, 40, 39].

Women's performance was not significantly different from men's because women were found to often have more efficient learning behaviors, which outweighs the statistical anxiety, leveling the performance according to Macher's 2012 study of 147 social science undergraduate students [87]. Instruction method has shown to impact anxiety and performance as well, Ghani and Maat's 2018 literature review found that courses with computer-based statistics learning resulted in reduced statistical anxiety and increases statistical performance and had a positive impact on the success of the course [53], highlighting the positive effect technology and software can have on statistics-related outcomes.

2.1.7 Comprehension and Statistical Anxiety

Difficulties in interpreting and reporting inferential statistics are not limited to students or beginners; even those with extensive statistical knowledge encounter these challenges [83]. Bakker and Wicherts' 2011 review of 281 psychology journal articles found that between 35% and 55% contained incorrectly reported p-values, 18% reported erroneous statistical results, and approximately 15% included at least one unsupported statistical conclusion, and as many as 4% of articles reported incomplete statistical findings [2].

The problem may even make it through to the peer-review process, Hardwicke and Goodman (2020) examined 107 biomedical journals to see how many used professional statisticians to conduct peer-review on statistics related works, they revealed that 34% the journals rarely or never utilized specialized statistical review before publication [57]. This issue has remained largely unchanged since 1998 when 33% of journals similarly lacked dedicated statistical review [54], indicating little progress over nearly three decades. As empirical research expands across disciplines, reducing statistical misreporting and misinterpretation becomes

increasingly critical.

2.1.8 Readability and Statistical Anxiety

One factor contributing to interpretation errors may be the complexity of technical or numerical language, which can hinder both reading and writing processes. Collins and Onwuegbuzie (2007) examined the impact of readability on statistical anxiety in a sample of 92 graduate students and found that lower reading comprehension was significantly associated with higher statistical anxiety levels [25]. While insightful, their study focused on graduate students, who typically have stronger reading skills than undergraduates. Since most undergraduate programs require at least one statistics course, students with less developed reading comprehension may experience even greater difficulty [105]. This challenge is likely further amplified for students with disabilities that affect reading ability [138]. The Flesch-Kincaid readability test offers a systematic way to evaluate the accessibility of written materials, and research shows that clearer statistical language and improved readability enhance comprehension of technical documents [71, 70]. Even among individuals with limited statistical training, greater readability in statistical output has been shown to improve both understanding and learning outcomes [94].

2.1.9 Disabilities and Statistical Anxiety

Although none of the five systematic reviews mention research involving statistics anxiety and disabilities, Lipka and Hess conducted a small qualitative study of those with learning disabilities ($n = 22$) while they were completing a specialized, closely-supported, online statistics course. Results showed that by the end the students felt more confident but did not change their perception of how difficult statistics tasks were. Their attitude towards statistics changed from worry and dread to a more affirmative outlook while still recognizing the tasks as difficult [81]. Research into impact of specific disabilities on statistics anxiety is virtually non-existent, with the few studies being primarily qualitative or interview based on very small sample sizes.

Regardless, the relationship between statistical anxiety and those with visual disabilities can be significant in statistics. Statistical anxiety may be exacerbated in this population due to the reliance on visual representations, charts, and graphs to understand or read [138] complex statistical information. Moreover, the lack of accessible statistical software or tools tailored to the needs of these individuals can further impede their ability to engage effectively with statistical tasks. Additionally, those with learning disabilities, such as dyscalculia or dyslexia, may encounter particular challenges when processing numerical information or understanding mathematical concepts can contribute to heightened levels of statistical anxiety, as individuals may feel overwhelmed or frustrated by the perceived complexity of statistical tasks.

2.1.10 Attitude Towards Statistics and Statistical Anxiety

A positive attitude toward statistics is consistently found to be indicative of success and performance in statistical tasks [19]. Individuals in math-intensive fields generally hold more favorable views of statistics compared to those in the social sciences and related disciplines [80, 11]. While extensive research has explored the link between mathematics anxiety and attitudes toward math, relatively little has been done to examine the distinct relationship between statistical anxiety and attitudes toward statistics. Studies suggest that negative attitudes toward statistics contribute to heightened statistical anxiety; however, existing definitions of statistical anxiety do not explicitly account for its connection to mathematics anxiety or attitudes toward statistics [5, 18].

2.1.11 Other Socio-demographic Characteristics and Statistical Anxiety

Some, bit relatively little, research has been done to study the differences in statistical anxiety between several other characteristics: Findings include international students showing higher anxiety than domestic students (though often very small samples) [101, 113], differences in anxiety between ethnicity groups (often comparing only Caucasian and African American participants and primarily conducted in only the US) [102] or cultures [4], and specific personality traits, like perfectionism, impacting levels of statistical anxiety [152, 92].

2.1.12 Software and Statistical Anxiety

Despite the extensive research on statistical anxiety, none of the validated statistical anxiety scales account for the role of statistical software in shaping users' anxiety or attitude. The most recent SAS 24-item scale does not mention statistical software [139], while the widely accepted STARS 51-item scale includes only a single item (item 12) about about inputting data into a computer, with no reference to statistical software use [29, 56, 22, 36]. A Chew and Dillon's 2014 review identified six validated statistical anxiety scales, none of which address statistical software directly, despite its central role in modern statistical analysis [18].

Some STEM-related fields have begun investigating how using software like R and SPSS in statistics courses affects student anxiety. While exposure and practice with these tools have been shown to reduce statistical anxiety over time, the anxiety remains prevalent, suggesting that software design and statistical output presentation may contribute to the issue [115, 111]. For instance, a study on biology students using R throughout a semester identified a combination of computer anxiety, programming anxiety, math anxiety, and statistical anxiety, collectively termed "R anxiety," which persisted throughout the course and was notably higher among women (2.1.2) [47]. Interestingly, Rafalski et al. (2019) found that prior coding and statistics experience did not significantly ease the learning curve for R, further emphasizing the challenges posed by statistical software [111]. However, Maxwell Hsu observed 207 business students in 2009 and found that increased statistical software self-efficacy promotes positive attitudes towards statistics, and anxiety has

the opposite effect [62].

Beyond the classroom, statistical software plays a crucial role in enabling researchers to create reproducible scripts, yet both students and professionals often struggle with effectively learning and using these tools for replicable research [146]. These findings suggest that statistical anxiety may be influenced not just by statistical concepts but also by the usability and design of statistical software itself. There are countless software and tools available, some coming with the promise of a support base and a large price tag (e.g. SPSS or SAS) and others being created and maintained by open-source contributors (e.g. R or Python), the usage of these tools varies among fields [89].

2.2 Literature Limitations

This literature review primarily focused on statistical anxiety, though related studies on math anxiety may provide additional insights. Additionally, potential connections between various measured factors, demographics, and statistical anxiety may exist but were not captured due to the review’s parameters.

While at least five major validated scales for measuring statistical anxiety exist (STARS[29], SAS[139], SAI[153], SAAS, SAM[36]) and are very thorough, they are often quite large ranging up to 51-items each. Using any of them multiple times in studies examining responses to statistical software during specific tasks would likely lead to participant fatigue and loss of valuable data due to limited task quantity. Moreover, these scales may not fully capture the nuances of statistics anxiety experienced in real-time, software-based statistical analysis. Though, perhaps a pre- and post- evaluation using these scales after learning and utilizing different software for an extended period of time may provide valuable insight into statistical software’s role or impact.

Research on statistical anxiety in relation to software use remains limited. Most existing studies focus on university classroom instruction, leaving a gap in understanding how statistical software influences anxiety in other educational settings, professional academia, or the workforce. No studies were identified that comprehensively examined both statistical anxiety and statistical software usage beyond higher education, highlighting an area for future research.

Chapter 3

Study 1: Does Learning Statistics Have to Be So Difficult? Evidence from a Randomized Controlled Trial

This chapter is based on a manuscript currently IN PRESS at ACM Transactions on Computing Education (TOCE). The first author is myself, Hannah Williams; Timothy Kluthe and Andreas Stefik contributed to the study design and writing. [150]

3.1 Introduction

Statistical reasoning plays a central role in modern education, research, and industry, influencing decisions across science, policy, business, and healthcare [16]. Despite its importance in many fields, statistical comprehension remains a widespread challenge, not only among students but also among professionals and the general public [2, 83]. These difficulties are exacerbated in computer science education. In our field, the use of complex methodologies and statistics is common, yet many institutions offer minimal training.

In the field of computer science, while programs vary, there is often a requirement to take at least one statistics course. There are many reasons, but a few might be its importance in data analysis, machine learning, or AI. Yet prior research consistently shows that such courses are among the most anxiety-inducing for students [86]. Given that adding more statistics courses to the undergraduate computer science education has pros and cons, an alternative is to design interventions that simplify the learning process. One promising idea, tested here for the first time, is Kluthe et al.’s recent “say what you do” [72] theory. Although statistical software has made data analysis more available and replicable; it may also introduce new obstacles: users often face steep learning curves, unintuitive code [72, 111, 44], and difficulties interpreting output [71, 101,

115, 83, 2].

The “say what you do” theory builds on the observation that traditional statistics naming conventions are often arbitrary or opaque “nomic mappings.” This term, while not used in the paper by Kluthe et al., was originally coined by Gaver in the context where certain sounds had no intrinsic meaning (e.g., a beep versus a raising pitch or speech) [52]. In statistics, test names like Tukey or Kruskal–Wallis are nomically mapped: they are people’s names and give no indication of what the test does. Thus, students have to memorize them, but human memory is not limitless [1]. Kluthe et al.’s theory proposes that using descriptive, non-nomically mapped names (e.g., Compare Means instead of ANOVA) could make statistics more approachable. However, the original study only examined participants’ preferences for names via a survey and did not assess the actual human impact on learning in more realistic contexts.

We expand on the “say what you do” theory with a full re-imagining and implementation of what statistics could be like. We use the names generated in prior work, but also add organization to the libraries to try and organize statistics thematically (e.g., tests that compare means, ranks, or medians). We modified the output to place the text at the eighth-grade reading level so it explains results in plain language, in addition to an “export to academics” mode with traditional APA reporting. Our approach is based on standard statistical equations under the hood, not LLMs. We compared its accuracy rigorously to R, Python, and Apache Commons (Java). We then implemented a wide variety of common statistical tests, from basic ANOVAs and descriptive statistics, to factor analysis and common corrections (e.g., Greenhouse-Geisser). Our approach hardly covers the entire field, but we were inspired by Sanders et al. in choosing what to implement [118]. Finally, we tested the theory in a randomized controlled trial.

During our study, we considered metrics like performance and also more subjective items, like how students felt while doing it. In reflecting on this question of user experience, we note that statistical anxiety is well-documented in educational research, with multiple validated instruments [29, 56, 139, 18]. Anxiety can impair performance, reduce engagement, and increase reliance on simplified tools. Existing scales, however, are lengthy and do not account for the kinds of software tools computer science students might use and tend to be around instruction or other matters. Participants completed tasks using either a traditional R-based system or an experimental prototype system developed for this study. Tasks spanned three phases: *selection* (choosing the appropriate test and understanding variable relationships), *input* (writing code), and *output* (interpreting results). Confidence and anxiety were assessed using definitions adapted from prior work on statistical self-efficacy and anxiety [101, 44].

Preliminary results largely support the “say what you do” theory: participants using the experimental system reported lower anxiety and higher confidence. Our expansion of the theory included simplified, plain-language output at an approximately eighth-grade reading level [71], alongside APA-style tables, which appeared to improve perceived clarity. This style of test names [72] and great care in organization and the automated written descriptions may have further reduced cognitive load and enhanced engagement. Thus, our results show it may be possible to significantly influence both the emotional and affective outcomes of

learning statistics.

We make three key contributions. First, we provide the first formal test of the “say what you do” theory in a randomized controlled trial, expanding the scope of prior work from terminology preference to a full evaluation of interface design effects on anxiety, confidence, uncertainty, and correctness. Second, we show that uncertainty can be measured separately from correctness. Participants who selected “I am not sure” may be expressing hesitation or disengagement rather than ignorance, a distinction that previous work conflates by classifying such responses as errors. Third, we examine these effects across a diverse sample including computer science and non-computer science students, different genders, and, to our knowledge for the first time in the literature, participants with disabilities. This allows us to compare our findings against prior work on field and gender differences while extending the evidence base to an underrepresented group.

An additional, smaller observation from this work is that during the implementation of our system, we encountered inconsistencies in the computations of widely used statistical software. For example, we identified issues with p -value behavior in the t -distribution implementation in Apache Commons Math under specific parameter settings (which has since been corrected), as well as issues related to ANOVA calculations in R’s *car* package for certain unbalanced designs. While fully evaluating the scope and impact of such issues is beyond the focus of the present work, these observations highlight the importance of transparency, interpretability, and trust in statistical software. This is particularly relevant in contexts such as computer science education, where statistical tools are frequently used but not always deeply examined.

3.2 Related Work

3.2.1 Statistical Anxiety: Scope and Evidence

Statistical anxiety, a feeling of fear or apprehension when confronted with statistical problems, data analysis, or interpretation tasks, has been consistently identified as a barrier to learning and performance from high school through professional practice [83, 68]. It impacts attitudes, performance, self-efficacy, and comprehension, and is particularly pronounced among students in non-mathematical fields. Large-scale estimates suggest that approximately 75–80% of graduate students report uncomfortable levels of statistical anxiety, with similar patterns in undergraduates; for example, Onwuegbuzie and Wilson [101] found that up to 80% of students enrolled in statistics courses report high anxiety levels, and Macher et al. [86] reported that students often rank statistics as the most anxiety-inducing course in their academic programs.

Seven major literature reviews have been done in the field. Onwuegbuzie and Wilson’s [101] review included 83 references, Baloglu and Zelhart’s [5] covered 72, Chew and Dillon’s [18] examined 56 studies (50 quantitative, six qualitative) and noted that samples were often overwhelmingly female (up to 100%) and predominantly from social sciences such as psychology and education. Macher et al. [85] included 11 studies on anxiety links with performance and other metrics, Ralston et al. [113] reviewed 81 references on pedagogy and demographic moderators, Ghani and Maat [53] synthesized 10 quantitative studies on

anxiety and achievement, and the most recent review by Trassi et al. [134] analyzed 199 articles, identifying 40 quantitative studies covering more than 11,000 participants, of which 67% were female, 80% were from human sciences, and had an average age of about 22 years. Across these reviews, common correlates of statistical anxiety include gender, age, academic field, course experience, and disability status, though the sample’s narrow composition, primarily university students from similar disciplines, makes it difficult to generalize the findings to other populations, such as professionals, older adults, those in different fields, or those with specific disabilities.

3.2.2 Measurement Gaps

Despite software’s central role in statistical work, validated statistical anxiety measures rarely address it. The widely used Statistical Anxiety Rating Scale (STARS; [29, 56, 22]) includes only one item about “putting data into a computer” (item 12) and three about seeking help with a “printout” (items 16, 19, 23), which may or may not reference software output. The Statistical Anxiety Scale (SAS; [139]) omits software entirely, with one item about understanding a printout (item 17). Chew and Dillon’s [18] review lists six validated scales (including STARS, SAS, SAI [153], SAAS, and SAM [36]) and none explicitly measure anxiety arising from statistical software use. This is a significant omission given that most modern statistical work is mediated through software, where usability, terminology, and output presentation can shape both understanding and affective responses.

3.2.3 Software-Specific Studies

Only a small body of empirical research links software use to statistical anxiety, confidence, or engagement. Forrester et al. [47] studied 215 undergraduate ecology students learning R and found that reduced disengagement and avoidance predicted increased self-concept and lower anxiety, with women showing more disengagement. Rafalski et al. [111] examined 154 students from multiple fields and found that prior statistics experience did not improve ability to learn R, suggesting software-specific learning challenges. Hsu et al. [62], in a study of 207 business students, reported that higher statistical software self-efficacy predicted more positive attitudes toward statistics. Rode [115] and Masuadi [89] highlighted differences in tool adoption (e.g., R, SPSS, SAS, Python, Excel) across fields, with potential affective implications, though direct anxiety measures are often missing. Widder et al. [146] noted that both students and professionals struggle to use statistical software effectively for reproducible research, again underscoring the role of tool design in shaping user experience.

3.2.4 Demographics and Anxiety Interaction

Gender differences in statistical anxiety are among the most consistently reported findings in the literature [11, 47, 37, 20, 90], though not all studies agree. For example, Mji [95] found no gender gap among

226 technology students (mostly women), and Kian et al. [68] reported lower anxiety among postgraduate women in a small qualitative study (N=36). Men may experience greater anxiety when seeking help [87, 84], and field composition plays a role: Trassi et al. [134] found that over 80% of statistical anxiety study samples came from human science fields, while only 15% came from STEM. Age effects are mixed, Baloglu [3] reported higher anxiety in older women, while Ralston et al. [113] observed higher anxiety in older men.

While gender differences in general computer and programming anxiety have been documented [114], the extent to which they apply to statistical software remains unknown in part because validated measures omit the impact of software. Software imposes unique demands, specialized terminology, complex numerical output, and code-based workflows for reproduction that could interact differently with demographic factors than do general computing tasks. Without targeted measures, it is unclear whether demographic patterns observed in statistical anxiety generalize to software contexts, let alone specific contexts like computer science.

3.2.5 Direct Precursor: Terminology in Statistical Software

The most directly related prior work is a randomized controlled trial with 118 computer science students that tested whether altering statistical function names in a programming environment could improve user preferences [72]. That study compared conventional statistical terminology (e.g., “ANOVA,” “LeveneTest”) to descriptive, “say what you do” names (e.g., “CompareMeans,” “CompareVariances”). In this approach, statistical tests were renamed to reflect their functional purpose, sometimes grouping procedures into higher-level conceptual categories. This structure is consistent with broader trends in statistics and quantitative methods education that emphasize thematic, concept-driven organization of statistical reasoning (e.g., relationships, differences, and prediction) rather than strictly technical or nomenclatural instruction [151]. The intervention revealed a significant increase in preference for the descriptive names ($p < .001, \eta^2 = .247$), suggesting that even small changes in terminology can influence how users engage with statistical tools.

The present study extends this prior work in two important ways. First, where Kluthe et al. examined only participants’ preferences for renamed statistical functions in isolation, we evaluate the downstream effects of those naming changes on anxiety, confidence, uncertainty, and correctness during realistic task-based contexts. Second, we expand the scope of design modifications well beyond terminology, incorporating restructured code syntax, plain-language output at an accessible reading level, and thematic organization of statistical procedures, and test these changes together as a coherent interface redesign to align with the “say what you do” approach. This moves the inquiry from asking whether users prefer descriptive names to asking whether a statistical environment meaningfully changes how users feel and perform when doing statistics. Our environment here is primitive and has limits, but thus expands on prior work meaningfully.

3.2.6 Summary

Statistical anxiety is a common and well-documented phenomenon, affecting up to 80% of students and often making statistics the most anxiety-inducing required course [101, 18, 134]. Demographic factors influence

anxiety levels: women, older students, undergraduates, students in human science fields, individuals with lower reading comprehension, international students, and some ethnic groups report higher anxiety [113, 53, 44, 37, 90, 138, 93, 4, 95]. Prior statistics course experience can improve comprehension and attitudes but does not eliminate anxiety [85, 134].

Higher statistical anxiety is associated with lower confidence and self-efficacy [101, 44, 90], poorer performance [85, 113], more negative attitudes toward statistics [134, 108], suboptimal learning behaviors such as procrastination and academic dishonesty [40, 119], and avoidance or disengagement [18, 85, 47]. Anxiety may also lead to misinterpretation, under-comprehension or reliance on others [105, 87], and reduced retention in statistics-related fields [113, 4].

Despite a robust literature on demographics, confidence, and performance, the impact of statistical software design, terminology, usability, and accessibility, remains largely unexplored. The few studies that include software suggest it can influence confidence, engagement, and perhaps anxiety, but the evidence base is small, methodologically diverse, and often indirect. This gap leaves critical questions open at the intersection of Human-Computer Interaction (HCI) and statistics computing education.

3.3 Research Question

Building on gaps in the literature, the primary goal of this study is to examine whether and how statistical software interfaces influence user experience and task performance. Specifically, we ask:

Can intentional changes to software interface design affect human confidence, statistical anxiety, perceived task difficulty, or task performance (correctness and uncertainty) when completing statistical tasks?

We investigate these outcomes in the context of two different statistical software designs: *Experimental*, a prototype “say-what-you-do” interface built using a prototype statistics programming language, drawing on recommendations from prior work on statistical nomenclature [72], and *Traditional*, a standard script-based, open-source statistical environment (R). The interface conditions were designed to explore how design decisions influence statistical task engagement, especially among novice users.

Gender was included as a fixed effect, informed by research indicating gender-based differences in statistical anxiety and self-efficacy [101, 114]. While gender-based differences in statistical anxiety and confidence are consistently observed, performance differences are generally small or absent, suggesting that any differences in accuracy would likely be small [12, 87]. Further, the cognitive complexity and ambiguity of statistical software has been linked to increased anxiety and reduced confidence [18], motivating the present software design intervention and its potential to mitigate such challenges.

In addition to the primary research question, we examined correlations among anxiety, confidence, and uncertainty, and compared perceived task difficulty across the three task domains (selection, input, and output) to validate prior findings and provide broader context for the role of design in statistical task engagement. As part of this, we try to capture whether participants felt uncertain during the study about

their choices, allowing a response where they can indicate that they do not know the answer to a question using “I am not sure”.

3.4 Method

3.4.1 Inclusion and Exclusion

Participants were recruited from a wide range of university-based sources, including departments in which statistics is a required component of the curriculum, disability and accessibility resource centers, and postings in graduate student centers and libraries, to ensure broad but domain-relevant participation. To be eligible for participation, individuals had to be at least 18 years old in accordance with consent requirements. Two undergraduate students reported their ages as 15 and 17; whether these reports were mistaken or not, both cases were excluded. For gender demographics, participants selected from four response options: Male, Female, Other, and Prefer not to say. Due to small sample sizes in the Other and Prefer not to say categories, these participants were excluded from quantitative analyses as it was not enough information to quantify impacts for the group. However, their responses were retained for qualitative analyses where applicable. To ensure data integrity, participants reporting implausible ages (under 10 years or over 100 years) were excluded, as age was a covariate in the study. Participants who did not complete the full study were included in analyses up to the point where responses were available. Analyses requiring complete data excluded participants with missing responses to maintain consistency and validity.

3.4.2 Participant Characteristics

A total of 320 participants began the study, with 297 completing at least the first task. Among these, four participants selected “Other” or “Prefer not to say” in response to the gender question. Among the remaining 293 participants, 197 were men and 96 were women. The Experimental group included 147 participants (97 men, 50 women), while the Traditional group consisted of 146 participants (100 men, 46 women), as shown in Table 3.1.

Participants reported a range of educational backgrounds. Specifically, 106 had some college or university experience, 72 had completed high school or a GED, 39 held an associate’s degree, 44 held a bachelor’s degree, 23 had completed a master’s degree, 7 held a professional doctorate, and 2 had earned a Ph.D. Participants also reported their current enrollment status: 107 were undergraduate seniors, 83 juniors, 36 sophomores, and 24 freshmen. At the graduate level, 38 were enrolled in graduate programs and 12 were post-graduate students. Eighteen participants were not currently enrolled or reported that enrollment was not applicable, and 3 were non-degree-seeking. This education distribution reflects a spectrum of learners at different stages of engagement with statistics, including current students, former students, and individuals applying prior training outside of formal academic settings. Such diversity is consistent with the range of contexts in which statistical tools are encountered and supports evaluation across varying levels of prior knowledge.

Table 3.1: Participant distribution

Software Design	Gender	Demographics	First Task	Last Task
Experimental	Female	52	50	47
	Male	105	97	87
	Other	1	1	1
	Prefer not to say	1	1	1
Traditional	Female	51	46	41
	Male	107	100	91
	Other	2	1	0
	Prefer not to say	1	1	1
Total		320	297	269

Participants reported a broad range of academic majors, which were grouped into broader fields of study to support analysis of disciplinary differences. A majority (158 participants, 54%) came from the formal sciences, of which 146 reported Computer Science as their major. Other fields included business and economics (34), engineering and technology (30), social sciences (30), and health sciences (20). The remaining 21 participants reported majors in education, the arts and humanities, natural sciences, or indicated no specific academic field. The diversity in broader disciplinary background was intentionally retained, as prior work suggests that statistics anxiety and performance differ across fields, particularly between quantitatively intensive disciplines and others. Capturing this variation enables evaluation of whether observed effects generalize across learners with differing levels of quantitative training. This is particularly important for computing education contexts, where statistical tools must support learners from varying disciplines.

Twenty-one primary languages were reported. English was the most common, identified by 231 participants (79%). Other frequently reported primary languages included Spanish (15), Chinese (8), Korean (5), and French (4). Additionally, Tagalog, Persian, and Hindi were each reported by 3 participants, Vietnamese reported by 2, and 9 participants reported other languages.

Participant ages ranged from 18 to 74 years ($M = 24.7$, $SD = 7.31$) and the average age within each group was between 24 and 25 years. Participants also reported their experience with statistics courses ($M = 1.00$, $SD = 1.42$). Random assignment was used for the design factor so the impact is due partially to chance, but the men in the Experimental group had the lowest statistical experience ($M = 0.84$) and the women in the Traditional group had the highest ($M = 1.09$). Among the 293 included participants, 111 (37.9%) reported no prior coursework in statistics. Another 122 (41.6%) had completed one course. Forty-four participants (15.0%) reported having completed two courses, and a small number had taken three or more (6 had taken three, 4 had taken four, and 6 had completed five or more). The most experienced participants reported completing 11 and 14 courses, respectively.

Of the 293 analyzed participants, 64 (21.8%) reported one or more disabilities. The most frequently reported condition was attention-deficit/hyperactivity disorder (ADHD), affecting 26 participants. Fifteen reported a mental health condition, 8 reported autism, and 8 reported a visual impairment, including 3 participants who were blind. Five were hard of hearing and 2 reported deafness. Other conditions included chronic health-related disabilities (8), learning disabilities (3), mobility impairments (2), and speech-related impairments (1). Six participants reported other disabilities not listed in the provided categories. Several participants reported multiple co-occurring disabilities. To our knowledge, our study is the first to report effects on those with disabilities in regard to our outcome measures.

To support statistical analysis and enhance interpretability, some participant characteristics were re-categorized prior to analysis. For instance, academic majors were collapsed into two broader groups (e.g., formal sciences and other) to ensure sufficient and meaningful representation across conditions. The justifications for these re-categorizations are detailed in Section 3.4.12.

Non-student subgroup A small subset of participants mentioned above ($n = 21$; 18 of which completed at least the first task) were not currently enrolled in an academic program or reported non-degree-seeking status. This group ultimately consisted of individuals with prior higher education experience, including bachelor’s ($n = 6$), master’s ($n = 6$), doctorate ($n = 4$), associate’s ($n = 1$), and some college education ($n = 1$). These participants were on average older ($M = 38$ years) and reported more prior exposure to statistics ($M = 4$ years of experience; $M = 2$ statistics courses) than the overall sample, suggesting that they were former students rather than individuals without academic training. Furthermore, this non/former student subgroup was small and slightly uneven across conditions (Experimental: $n = 8$; Traditional: $n = 10$). Given this limited sample, we do not conduct inferential analyses specific to this subgroup. However, their demographic and educational profiles substantially overlap with the study’s age and education covariates (e.g., undergraduate vs. graduate classification and age grouping discussed in Section 3.4.12), suggesting that current or former student grouping would not introduce a qualitatively distinct participant category beyond what is already modeled in the primary analyses. Accordingly, results should be interpreted as reflecting adult learners spanning both current students and individuals with prior academic training, rather than exclusively enrolled students.

3.4.3 Sampling Procedures

Responses were collected over a six-month period between May and December 2024. Participants self-selected into the study via an online platform hosted on a secure server. Participants were directed that they could not use a mobile device due to the coding portion of the study. Upon trying to access the site, if a mobile device was detected, entry into the study was disabled. To ensure balanced comparison groups by gender, a fixed factor in the study, participants were randomly assigned to one of two software conditions based on their response to the gender question.

Recruitment involved outreach to hundreds of professors across more than thirty departments and fields, as well as to disability groups and resource centers affiliated with two universities. In-person recruitment also occurred through instructor-led classroom announcements and outreach events at the campus libraries. Of the individuals who accessed the study ($N = 354$), 322 completed the demographics survey, and 299 completed the first task. Thus, approximately 85% of those who accessed the study provided sufficient data to be included in at least some analyses.

Where applicable, participants were offered 2% extra credit contingent on instructor approval, with alternative options provided for those who declined to participate. Upon study completion, instructors who had approved the extra credit option received a list of student participants who had opted in using their university email addresses; participant data were anonymized immediately after this step.

This study was approved by the authors' Institutional Review Board and conducted in accordance with ethical guidelines for informed consent, confidentiality, and voluntary participation.

3.4.4 Sample Size, Power, and Precision

As software has often been left out of the major studies on statistical anxiety, we undertook a set of pilots to help us sort out issues with samples and design. Notably, four pilot studies were conducted to refine the study's questions, tasks, and scales, with a focus on clarity and appropriate difficulty. The first three pilots, run on the Qualtrics platform, compared different test nomenclature and output formats: English-based, R-style output, and experimental chart-based output. These involved 3, 16, and 54 participants, respectively. They used a two-item anxiety scale with opposing statements, following Rode et al. [115]. However, the third pilot revealed an issue: many participants endorsed both opposing statements (e.g., "I do not feel any anxiety..." and "I feel very anxious..."), indicating they did not recognize the contradiction and thus compromising the scale's validity.

A fourth pilot ($N = 20$), hosted on a university server, focused on qualitative feedback about study structure and design. Switching from the Qualtrics platform to a university server provided us the ability to customize questions and measurement types. This pilot included three between-subjects conditions (two experimental designs and one R-design), closely resembling the final study. Although it retained the two-item anxiety scale, bolded language and explicit reminders (e.g., "please read carefully") were added to clarify the opposition. While this improved understanding slightly, confusion persisted. As a result, the final study adopted single-item measures for both overall anxiety and overall confidence and included sub-scale items to aid in identifying key aspects contributing to each. Feedback from the fourth pilot also led to the inclusion of clear definitions for "anxiety" and "confidence," revisions to question wording for clarity, and adjustments to programming task difficulty to more effectively contrast the software designs while maintaining similar task durations (e.g., simple t-tests alone were insufficient). Because many pilot participants reported that the tasks felt like a test and that they sought external help, the final study included a consistent reminder at the top of each task: "You are not being tested. Do not seek any outside assistance. Simply answer to

the best of your ability.” The pilot data was not directly comparable to the final dataset due to iterative design changes.

Participants from pilot studies were excluded. In the final sample, new participants were recruited and approximately 50 participants were retained in each female group and 100 in each male group who completed at least the first task (see Table 3.1).

3.4.5 Measures and Covariates

After an extensive piloting period where we evaluated how to study the software hypotheses we were interested in and trying to estimate the kinds of samples that would be necessary, we finalized our measures and proceeded to the primary study. This study collected several measures across six tasks designed to assess participants’ understanding of common statistical methods and their experiences engaging with these tasks.

3.4.5.1 Primary Measures

These core measures aligned with the study’s research questions and hypotheses. They were collected consistently across all tasks and formed the foundation of the main analyses.

- **Anxiety.** Existing statistical anxiety scales [29, 56, 139, 36, 153], including the widely used 51-item STARS capture a range of contributing factors to statistical anxiety. However, they do not include items explicitly related to software. Administering up to 51 items for each of the six tasks in this study would likely lead to participant fatigue and reduce the quality of responses. Thus, based on pilot testing, the present study instead employed a single-item, 7-point radio-button Likert scale (“Please indicate how much anxiety you felt while completing this task,” from *none at all* (1) to *an extreme amount* (7)) to measure overall task anxiety. Before each rating, participants were reminded of the intended definition of anxiety as “*a feeling of fear or apprehension when presented with statistics content and problems or when gathering, processing, and interpreting data.*” In addition, each task area included tailored subscales that asked, “*How did the following contribute to any anxiety you felt?*” using the same 7-point range. For selection tasks, contributors included: data format; explanation of the data; determining variable types; selecting the most appropriate statistical test; question complexity; statistical test naming conventions; and familiarity with terms or concepts. For input tasks, contributors included: data format; explanation of the data; code format; explanation of the task; example code; amount of code; naming conventions in code; code complexity; and familiarity with code, terms, or concepts. For output tasks, contributors included: output format; amount of output; output complexity; numerical values; tables (if any); question complexity; code used to produce the output; and familiarity with terms or concepts.
- **Confidence (Self-Efficacy).** Finney and Schraw’s CSSE and SELS scales [44], consisting of 14 items each, measure statistical self-efficacy, but their items focus on abstract or general knowledge

(e.g., “Explain the difference between a sampling and a population distribution”), making them less suitable and likely too cumbersome for capturing confidence in real-time statistical software use. In this study, we adopted the same terminology from Finney and Schraw’s scales, as such “confidence” refers to task-specific statistical self-efficacy. Confidence was assessed with the same single-item, 7-point radio-button Likert format used for anxiety, replacing the term “anxiety” with “confidence.” Before each rating, participants were reminded of the intended definition of confidence in this context as *“the sense of one’s own ability to understand and complete actions required to accomplish specific statistics-related tasks.”* As with anxiety, each task type (selection, input, and output) included the same tailored sub-scales described above, asking how each listed aspect contributed to participants’ confidence levels.

- **Uncertainty Rate.** Participants had the option to respond “I am not sure” to any question or task. This metric captures the frequency of that response, reflecting their self-perceived uncertainty. Patterns of uncertainty were assessed to gauge signs of disengagement.
- **Correctness Rate (Accuracy).** For each task, correctness was determined by mathematical accuracy, code precision, and alignment with widely accepted interpretations of statistical tests and outputs. To better reflect engaged responses, answers of “I am not sure” were excluded from scoring prior to calculating correctness rates, thereby reducing the influence of uncertainty or disengagement on performance measures.

3.4.5.2 Secondary Measures

Secondary measures offered additional context for interpreting results but were not directly tied to the study’s hypotheses. Accordingly, they were interpreted with caution.

- **Completion Rate.** Retention was tracked across the study, starting at 100% immediately after the demographics phase and in some cases decreasing with each subsequent task. Patterns of retention were tracked.
- **Time.** Task completion time was measured from the moment the page loaded until the participant submitted their response.
- **Difficulty.** During the final feedback phase, perceived difficulty for each task area was rated on a single-item 5-point Likert scale ranging from “very easy” (1) to “very difficult” (5). This anchor set is based on Vagias’ (2006) compilation of Likert-type response options, which has been widely adopted in survey methodology and practice [136]. Unlike the repeated 7-point measures of anxiety and confidence, this difficulty rating was collected only once at the end of the study as a retrospective summary and was not directly analyzed alongside the affective scales.

- **Feedback.** Open-ended responses were collected for each task to capture participant perspectives on anxiety and confidence, and collected in the final feedback phase for anxiety, difficulty and accessibility.

3.4.5.3 Covariates

The covariates were included in analyses to control for individual differences that could influence the relationship between software design and participants' outcomes.

- **Age.** Self-reported in years.
- **Primary Language.** Self-reported language spoken at home.
- **Education Level.** Derived by combining current enrollment year with highest completed education to yield a composite measure.
- **Academic Field.** Participants' majors were grouped into broader disciplinary categories.
- **Statistics Courses Taken.** Self-reported count of completed statistics courses.
- **Disability Status.** Participants indicated whether they had a disability and were able to provide more details via a selection box or open-text form.

3.4.5.4 Unreported Measures and Covariates

The following variables were collected but excluded from analysis due to redundancy, irrelevance to the study's aims, or limited variability within the sample:

- **Statistics Job Experience.** This variable was reported by few participants and showed limited variability. In several cases, participants reported more years of statistical job experience than their age, indicating inconsistencies. As such, this measure was deemed unreliable and excluded from analyses.
- **Institution.** Most participants attended the same institution, or listed no institution, limiting comparisons.
- **State/Country.** This was deemed irrelevant for core analyses and lacked sufficient representation for geographic comparisons.
- **Highest Education Level Completed/Education Year.** These were combined into a single derived variable, making individual categories redundant.
- **Education Major.** Original open-ended responses were condensed into broader academic field categories for analysis.

3.4.6 Data Collection

The tasks were embedded in an online survey hosted on a server and accessed via a secure link. To ensure data integrity and prevent duplicate entries, the study employed both session tracking and email validation. Session tracking restricted users from submitting multiple entries from the same browser session, while email validation enforced a one-time participation rule, disallowing reuse of email addresses. These combined measures helped maintain the accuracy of the dataset by ensuring that each participant contributed only one response. Together, they mitigated risks of duplicate participation and artificial inflation of sample size. Following validation, extra credit was distributed to eligible students through their instructors, as approved by the IRB. All identifying information, including email addresses, was then removed to anonymize the dataset and protect participant confidentiality. Participants were assigned to either the Experimental or Traditional software condition based on responses to a gender demographic item (Male, Female, Other, Prefer not to say), aiming for a roughly equal gender distribution across design groups. Despite these efforts, some group imbalances persisted due to lower female participation and survey incompleteness in certain conditions.

3.4.7 Quality of Measurements

Throughout multiple pilot studies, participants were asked to provide feedback on their experience and flag any issues with the questions or tasks, which informed iterative improvements to subsequent versions. Several of the scales used in this study were single-item measures rather than established multi-item instruments. As such, these scales are experimental in nature, relying on participants' self-assessment of their own levels of anxiety and confidence in response to each question or task. To support the reliability and validity of these measures, we conducted test-retest reliability analyses on the anxiety and confidence items independently for each task area. Additionally, we examined correlations between anxiety and confidence ratings across tasks and aspects, as prior literature consistently reports a negative correlation between these constructs [108, 134, 87, 66].

As a further validity check, we examined anxiety and confidence ratings for items that should not differ between the design groups. For example, tasks involving the identification of variable types (categorical, numerical, ordinal) presented identical information across groups, and thus any differences would be unexpected. Finally, to ensure inter-rater reliability in grading for the input area of the study, we developed a structured grading manual and scoring rubric. Two independent raters applied these criteria to a sample of responses (15 per software group, ~10% of total responses), and Cohen's Kappa was used to evaluate agreement. Following validation, the remaining responses were graded by a single rater. Further details on the grading process are provided in the Psychometrics section 3.4.10.

Table 3.2: Overview of the selection tasks. See Appendix A.1 for full details.

Task	Description
1	Participants were introduced to a fictional dataset of apples with variables including weight, tree height, color, and taste. The first five questions addressed data classification, and participants then identified whether the first row referred to the same apple across columns. They also selected the correct statistical test for two relationships: Weight vs. Tree Height (continuous) and Color vs. Taste (categorical). (See Appendix A.1, Figures A.1 and A.2.)
2	This task used a dataset with heart rate measurements for four species. Participants classified variables, identified independent and dependent variables, and determined if measurements were repeated. They then selected the appropriate statistical test (one-way ANOVA). (See Appendix A.1, Figures A.3 and A.4.)

3.4.8 Instrumentation

3.4.8.1 Tasks

Participants completed six tasks covering fundamental statistical concepts and analyses, grouped into three domains: test selection (multiple choice), code implementation (open-ended), and output interpretation (multiple choice). Figures showing the full tasks as presented to participants are provided in Appendix A.1, including both Experimental and Traditional versions, as well as the correct answers.

Question 6: Which statistical test would be most appropriate to determine if the average bpm is different between the four species?

- ☐ CorrelateGroups()
- ☐ CompareMeans()
- ☐ CompareVariances()
- ☐ CompareCounts()
- ☐ I am not sure
- ☐ There is not enough information

Figure 3.1: Experimental task 2 - question 6

3.4.8.2 Feedback

At the bottom of each task, participants rated their anxiety and confidence using a 7-point Likert scale and were able to provide open-ended information and comments regarding the ratings they gave. Additionally, sub-scale ratings captured specific factors that contributed to their anxiety or confidence. At the end of the study, participants rated how difficult they perceived each task area to be on a 5-point Likert scale and had

Question 6: Which statistical test would be most appropriate to determine if the average bpm is different between the four species?

- ☐ cor.test()
- ☐ aov()
- ☐ t.test()
- ☐ chisq.test()
- ☐ I am not sure
- ☐ There is not enough information

Figure 3.2: Traditional task 2 - question 6

Table 3.3: Overview of the input tasks. See Appendix A.1 for full details.

Task	Description
3	Participants implemented a one-way ANOVA in code, modifying an example script with dataset and variable names. Task requirements differed slightly by software design. Correctness was manually evaluated based on logical accuracy, syntax, and semantic alignment. (See Appendix A.1, Figures A.5 and A.6.)
4	This task required a repeated-measures ANOVA using a medication dataset with three time points. Participants modified the dataset load function and variable names according to their software design. Correctness was evaluated as in Task 3. (See Appendix A.1, Figures A.7 and A.8.)

the opportunity to provide open-ended feedback regarding task difficulty, accessibility issues, and any other relevant experiences. The scales and sub-scales are described in detail in Section 3.4.5.

3.4.9 Masking

Participants were not informed that the study involved two different software design conditions, ensuring that their responses were not influenced by awareness of group assignment. Pseudo-random assignment was performed automatically based on demographic information, and researchers did not manually control group placement. During analysis, group assignments and demographic variables were included in statistical models, but no individual participant data was examined in a way that could introduce bias. This maintained a degree of partial blinding throughout the analysis process.

3.4.10 Psychometrics

This section outlines the psychometric properties of the measures used in this study, including inter-rater reliability for coding evaluations, test-retest reliability for anxiety and confidence ratings, and the relationship between anxiety and confidence.

```

DataFrame frame
frame:Load("example.csv")
frame:SelectInputFactor("groups")
frame:SelectOutputVariable("values")
frame:TestAssumptions(true)
frame:CompareMeans()

```

Figure 3.3: Experimental task 3 - example code

```

frame = read.csv(file="example.csv")
frame$groups = as.factor(frame$groups)
aov_result = aov(values~groups, data=frame)
shapiro.test(aov_result$residuals)
leveneTest(aov_result)
summary(aov_result)
effectsize::eta_squared(aov_result)

```

Figure 3.4: Traditional task 3 - example code

3.4.10.1 Code Evaluation and Inter-Rater Reliability

Inter-rater reliability for code grading was assessed to ensure evaluation consistency. A grading manual, similar to the token-based approach by Daleiden et al. (2020) [30] but incorporating both syntax and semantic features, was developed to address the substantial differences between the two programming languages used.

Two independent raters scored a randomly selected 10% subset of responses for Task 3 (Table 3.5) and Task 4 (Table 3.6), assigning a score of one if a criterion was met and zero otherwise. Cohen's Kappa was calculated ($\kappa = .899, SE = .018$) to determine inter-rater agreement, with results reported in the Results section. Correctness rates for programming tasks were calculated using this grading system. For example, a participant receiving a zero on criterion 12 of Task 3 but a one on all other criteria would receive a score of 91.66% for that task.

Due to programming language differences, some criteria were not directly comparable between groups. In Task 4, the Traditional group had one additional criterion related to the subject identifier in repeated measures. Consequently, missing one criterion resulted in a score of 94.12% for the Traditional group and 93.75% for the Experimental group. This slight imbalance was negligible for the analysis in practice, but we point it out because any such analysis involving more than one programming language does, by definition, have a non-exact mapping between them (otherwise it would be the same language).

3.4.10.2 Creating Single-Item Measures of Anxiety and Confidence

A key methodological challenge in this study was measuring statistical anxiety and confidence in a way that supports repeated, task-level assessment. Established instruments such as the Statistical Anxiety Rating

Table 3.4: Overview of the output tasks. See Appendix A.1 for full details.

Task	Description
5	Participants interpreted the output of a t-test, including assumption checks, group statistics, and effect size. Four comprehension questions assessed their understanding of normality, sample sizes, significance, and effect size. (See Appendix A.1, Figures A.9 and A.10.)
6	Participants analyzed the output of a one-way ANOVA, including assumption checks, main results, eta-squared effect size, and Tukey post hoc comparisons. Five comprehension questions evaluated understanding of assumptions, significance, effect size, and group differences. (See Appendix A.1, Figures A.11 and A.12.)

Scale (STARS; 51 items) and other commonly used measures (e.g., SAS [139], SAI [153], SAAS, SAM [36]; ranging from 20–51 items) are designed to capture broad, stable perceptions across instructional or evaluative contexts. Similarly, statistical self-efficacy scales (e.g., Finney and Schraw) include 14 items each. Administering these instruments after each of the six tasks in this study would require hundreds of item responses per participant, introducing substantial response burden, likely exceeding what an institutional review board would consider reasonable. In addition, these measures primarily reflect reactions to coursework, exams, or general exposure to statistics, rather than moment-to-moment interaction with statistical software and tasks. As such, one immediate issue in a study like ours is how to make quick, clear, assessments of anxiety or confidence.

In pilot studies, the most promising approach we tried was from Rode et al. [115]. This measure was intended to capture context-specific anxiety when using R or SPSS software: “I do not feel any anxiety when I look at this statistical output” (reverse-coded) and “I feel very anxious when I look at this statistical output,” rated on a 7-point scale. Across pilot studies ($N = 54$ and $N = 20$), many participants endorsed both items simultaneously, indicating that the contradiction between the statements was not reliably recognized. This issue persisted despite revisions to instructions and visual emphasis. Prior research has shown that reverse-worded items increase cognitive processing demands and are more susceptible to inattentive responding, satisficing, and straightlining, often producing inconsistent or contradictory responses, particularly in repeated or low-engagement contexts [74, 131, 126, 130, 69]. We observed exactly these types of issues in our pilots. Given these constraints, we adopted a new approach, acknowledging that it carries its own limitations. We used single-item measures of anxiety and confidence grounded by definitions to guide context-specific responses. We next discuss the items and then how we evaluated them.

Statistical anxiety has been defined in various ways in prior literature, for example “a feeling of anxiety when taking a statistics course or doing statistical analysis” [29], “a form of state anxiety characterized by apprehension and fear when engaging with statistics-related content” [101], “the apprehension that occurs when an individual is exposed to statistics content or problems and instructional situations, or evaluative

Summary						
A comparison between 2 samples was tested. First, the samples were tested for normality and all samples appear to be normal. Results show that there may not be a difference between group1 and group2, $t(58.99) = 1.61$, $p = .112$, Cohen's $d = 0.41$.						
Table Summary						
comparison	statistic (t)	degrees of freedom (df)	probability (p)	significant (p < .05)	effect size (cohen's d)	effect size magnitude
group1_group2	1.61	58.99	0.112	no	0.41	small
Descriptive Table Summary						
sample	size (n)		mean (μ)		standard deviation (σ)	
group1	31		11.62		3.53	
group2	30		10.18		3.46	

Figure 3.5: Experimental task 5 - output

contexts that deal with statistics” [85]. There have been multiple refinements in subsequent literature that distinguish it from math anxiety [18]. As such, we provide a study-specific operational definition for this context that aligns with prior conceptualizations as “*Anxiety: a feeling of fear or apprehension when presented with statistics content and problems or when gathering, processing, and interpreting data.*”

In prior work, statistics self-efficacy has been defined as “confidence in one’s abilities to solve specific tasks related to statistics” and self-efficacy for learning statistics has been defined as “confidence in one’s abilities to learn the skills necessary to solve such tasks” [44], which situates these constructs within Bandura’s foundational work on self-efficacy theory [7]. In the current work, this was used as a study-specific, task-level construct: “*Confidence: the sense of one’s own ability to understand and complete actions required to accomplish specific statistics-related tasks.*” We want to be clear in saying definitions are not the same as validated scales. The question then became a balance of how to quickly spot check statistical anxiety or confidence and validate that we are observing something similar to prior literature.

To evaluate the alignment of the single-item measures with expected psychometric properties, we assessed reliability and construct validity using complementary approaches. First, we wanted to know if we took similar measures more than once if we would obtain the same answer. Thus, test-retest reliability was assessed via correlations between repeated ratings within each task domain. This allowed us to evaluate whether similar tasks would yield stable responses. Convergent validity was examined through the relationship between anxiety and confidence, which prior literature consistently reports as negative [87, 66, 108, 134]. Results of these evaluations are reported in Section 3.5.3, with full analyses provided in Appendix A.2.1.

```

Shapiro-Wilk normality test
data: group1
W = 0.968, p-value = 0.4659
-----
Shapiro-Wilk normality test
data: group2
W = 0.97853, p-value = 0.7852
-----
Welch Two Sample t-test data:
group1 and group2
t = 1.6112, df = 58.99, p-value = 0.1125
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
-0.3490012 3.2341625
sample estimates:
mean of x mean of y
11.62258 10.18000
-----
Cohen's d | 95% CI
-----
0.41 | [-0.10, 0.92]

- Estimated using pooled SD.Warning message:
Missing values detected. NAs dropped.
-----
group1 group2
11.62258 10.18000
-----
group1 group2
3.531450 3.461254
-----
group1 group2
31 30
-----

```

Figure 3.6: Traditional task 5 - output

Limitations are discussed in Section 3.7.5.

To provide succinct bottom line, we obtain similar results to prior work, including the classic negative correlation at a similar magnitude between statistical anxiety and confidence using this approach. However, our point is that since Rode et al. [115]’s 2-item approach did not replicate for us in pilot studies and others either did not apply or were far too bulky, we tried an alternative and this appeared promising. Even ignoring that statistical anxiety and correctness are only one part of this randomized controlled trial, how people feel about statistics is something we think we should consider as a community. We just want to be clear that we think the approach is promising in this context, but has limits that will be discussed.

3.4.11 Conditions and Design

This study employed a mixed-design randomized controlled trial, incorporating both between-subjects and within-subjects components and importantly included qualitative feedback from participants. The primary between-subjects factor was software design and a second fixed factor of gender, with repeated measures across tasks and observational data on other demographic factors. Participants were pseudo-randomly assigned to one of two conditions: the Traditional Software Design, using the default R programming environment for statistical analysis, or the Experimental Software Design, using a modified interface. Pseudo-randomization was stratified by gender to ensure balanced group sizes.

Participants completed two tasks in each of the three statistical domains. In the selection area, they

Table 3.5: Grading criteria for evaluating task 3

Experimental Criteria	Traditional Criteria
1. Data frame object is created	1. Data frame object is created
2. File name is correct	2. File name is correct
3. File name includes quotations	3. File name includes quotations
4. File name includes .csv	4. File name includes .csv
5. Input factor name is correct and includes quotations	5. Input variable is set as a factor correctly and selected/used correctly in the test parameters
6. Output variable name is correct and includes quotations	6. Output variable is selected/used correctly in the test parameters
7. Main test name and parameters are correct	7. Main test name and parameters are correct
8. Assumption/posthoc/effect size test name and parameters are correct	8. Assumption/posthoc/effect size test name and parameters are correct
9. There are no syntax mistakes	9. There are no syntax mistakes
10. There are no additional tests, factors, or variables	10. There are no additional tests, factors, or variables
11. There are no missing tests, factors, or variables	11. There are no missing tests, factors, or variables
12. Code appears as though it would produce the correct results	12. Code appears as though it would produce the correct results

identified the appropriate statistical test (Pearson correlation, chi-squared test, or ANOVA) based on provided datasets and scenarios. In the input area, they implemented statistical code for an ANOVA and a repeated measures ANOVA using new datasets. In the output area, they interpreted results from t-test and ANOVA output. Although participants completed two tasks in each domain, analyses were conducted using between-subject comparisons of averaged measurements within each domain, which accounts for the repeated measures effect. Primary dependent variables included statistical anxiety, confidence, performance, uncertainty, and perceived difficulty. Gender was treated as a fixed factor due to prior literature identifying differences in statistical anxiety and confidence [134]. Additional covariates included age, primary language, education level, prior statistics coursework, academic field, and disability status. This design enabled an evaluation of how interface design influences statistical anxiety, confidence, and performance while accounting for demographic factors.

3.4.12 Data Diagnostics

Data processing began with primary language, where 21 distinct responses were condensed into “English” and “Other” to distinguish native English speakers from those with other primary languages. This variable, while

Table 3.6: Grading criteria for evaluating task 4

Experimental Criteria	Traditional Criteria
1. Data frame object is created	1. Data frame object is created
2. File name is correct	2. File name is correct
3. File name includes quotations	3. File name includes quotations
4. File name includes .csv	4. File name includes .csv
5. Input factor name is correct and includes quotations	5. Output variable name (value) is set, includes quotations, and is correct in test parameters
6. Repeated factor name is set and includes quotations	6. Repeated factor name (key) is set and includes quotations
7. Repeated factor list of three levels is named correctly	7. Repeated factor list of three levels is named correctly
8. Repeated factor list of three levels is separated by commas and within one set of quotations	8. Repeated factor list of three levels is separated by commas and has separate quotations
9. Input variable is set as a factor correctly and selected/used correctly in the test parameters	9. Subject identifier variable is set as a factor correctly and selected/used correctly in the test parameters
10. Repeated variable is set as a factor correctly and selected/used correctly in the test parameters	10. Input variable is set as a factor correctly and selected/used correctly in the test parameters
11. Main test name and parameters are correct	11. Repeated variable is set as a factor correctly and selected/used correctly in the test parameters
12. Assumption/posthoc/effect size test name and parameters are correct	12. Main test name and parameters are correct
13. There are no syntax mistakes	13. Assumption/posthoc/effect size test name and parameters are correct
14. There are no additional tests, factors, or variables	14. There are no syntax mistakes
15. There are no missing tests, factors, or variables	15. There are no additional tests, factors, or variables
16. Code appears as though it would produce the correct results	16. There are no missing tests, factors, or variables
	17. Code appears as though it would produce the correct results

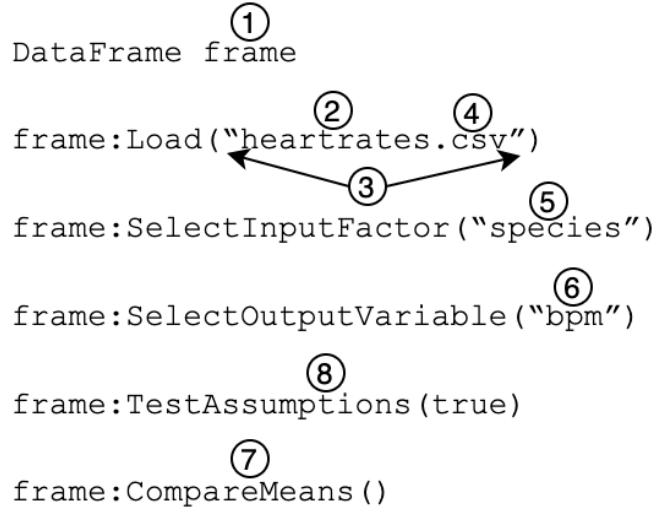


Figure 3.7: Experimental task 3 criteria. Numbers indicate which evaluation criterion (shown in Table 3.5) was applied to that portion of the code.

not a direct measure of international student status, serves as a proxy for language-related challenges that can heighten statistical anxiety [101], offering insight into potential comprehension and anxiety differences.

The random factor of majors, initially collected as open-ended responses with 125 variations (the full breakdown of each is discussed in section 3.4.2), were collapsed into “Formal Sciences” (e.g., Mathematics, Statistics, Computer Science, Data Science) versus “Other.” This approach reflects a theoretically grounded distinction between quantitatively intensive disciplines and those with less emphasis on formal mathematical training, which prior research has found may influence performance and anxiety [80]. The categorization also ensured sufficient statistical power and allowed academic field to be included as a covariate in subsequent analyses, enabling evaluation of whether the effects of statistical software design differ across levels of prior quantitative exposure.

Education level combined highest degree earned with current enrollment. Those enrolled in graduate school were classified as “Graduate” regardless of previous degree, while those with completed graduate degrees but not enrolled were initially labeled “Professional.” To simplify analysis and balance groups, “Graduate” and “Professional” were merged, yielding “Undergraduate” and “Graduate” categories that reflect both past and current academic engagement.

Statistics course experience, originally ranging from 0–14 courses, was unevenly distributed, with 37.9% having taken none and 41.6% only one. Responses were grouped as “None,” “One,” or “Multiple,” preserving meaningful distinctions between minimal and extensive exposure while balancing groups.

Age, initially continuous, was highly skewed, with only 47 participants over 30. Following prior work

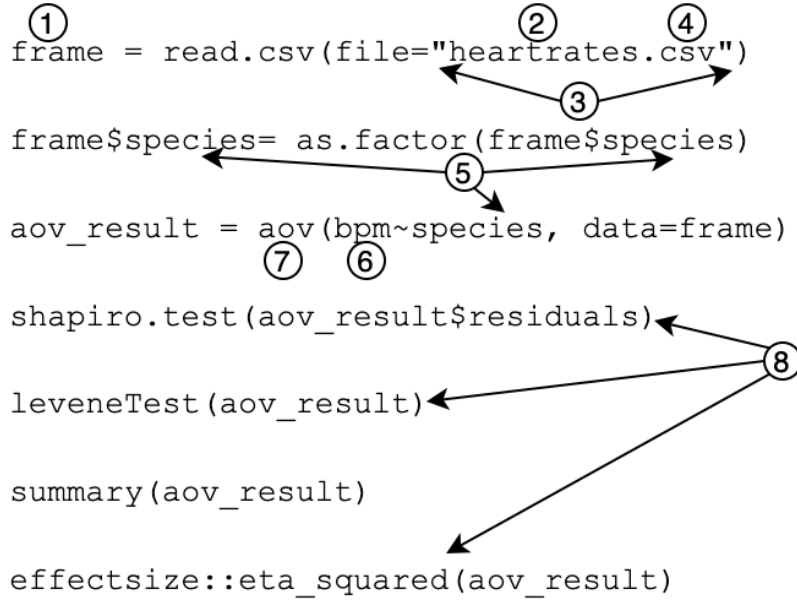


Figure 3.8: Traditional task 3 criteria. Numbers indicate which evaluation criterion (shown in Table 3.5) was applied to that portion of the code.

highlighting age differences in statistical anxiety, noting that the majority of research has involved those under 30 years old [37], participants were grouped as “Under 30” or “30 and over” to ensure balanced comparison and giving us a way to loosely compare to prior work.

Disability status included varied conditions with uneven representation across types, making separate analyses impractical. To maintain statistical feasibility while acknowledging potential impacts on anxiety, confidence, and performance, responses were classified as “Yes” (one or more disabilities) or “No.” We relied on qualitative feedback from these users especially to better understand their design needs.

In summary, the covariates included in the analysis were: disability status (“Yes”/“No”), education level (“Undergraduate”/“Graduate”), academic field (“Formal Sciences”/“Other”), primary language (“English”/“Other”), age group (“Under 30”/“30 and over”), and statistics course experience (“None”/“One”/“Multiple”). The representations of the categorized covariates in each group are shown in Table 3.7.

3.4.13 Analytic Strategy

To ensure robust results, multiple statistical methods were applied. Key measures like correctness with multiple choices were entirely objective and graded. Reliability and validity of anxiety and confidence were evaluated using Pearson correlations. Where applicable, especially code tasks, inter-rater reliability for coded responses was assessed with Cohen’s kappa.

Potential covariates (Primary Language, Academic Field, Statistics Course Experience, Age Group, Disability Status, and Education Level) were evaluated for redundancy using chi-squared tests with Cramer’s

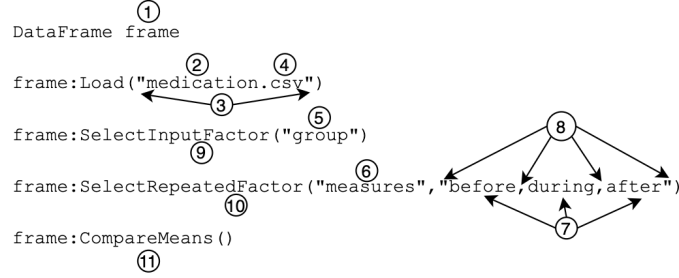


Figure 3.9: Experimental task 4 criteria. Numbers indicate which evaluation criterion (shown in Table 3.6) was applied to that portion of the code.

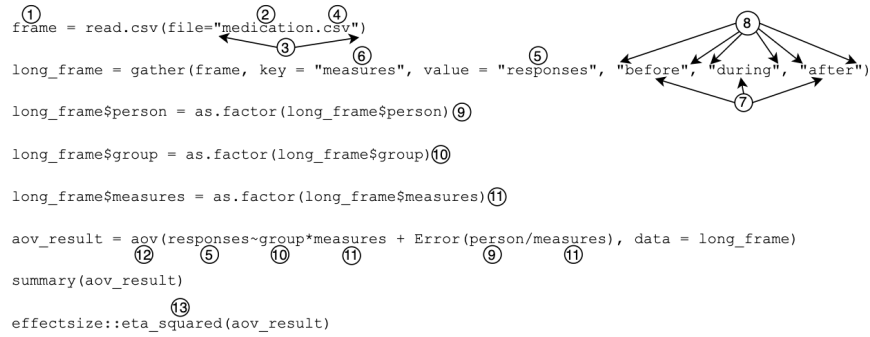


Figure 3.10: Traditional task 4 criteria. Numbers indicate which evaluation criterion (shown in Table 3.6) was applied to that portion of the code.

V as an effect size. Education Level was strongly associated with Academic Field ($p < .001$, $V = .25$), Statistics Course Experience ($p < .001$, $V = .33$), Age Group ($p < .001$, $V = .41$) and Primary Language ($p = .002$, $V = .18$), indicating considerable overlap. To avoid multicollinearity and sparse groupings, Education Level was excluded from subsequent models. The remaining covariates were retained individually, with Academic Field also modeled as an interaction with Gender due to a significant association between these factors ($p < .001$, $V = .28$).

Primary analyses used ANCOVA where applicable. Study-wide analyses were conducted first, followed by separate area-specific analyses (selection, input, output), with Tukey–Kramer post hoc tests where appropriate. In each area, “I am not sure” responses were analyzed separately from other responses to identify factors influencing correctness and uncertainty. In Input tasks, correctness was determined through rubric-based grading, while uncertainty was trichotomized due to the small number of questions, precluding ANCOVA

Table 3.7: Participant characteristics re-categorized

Characteristic	Category	Experimental		Traditional	
		Female	Male	Female	Male
Primary Language	English	42	81	35	73
	Other	8	16	11	27
Academic Field	Formal Sciences	16	58	16	68
	Other	34	39	30	32
Education Level	Graduate	12	19	11	18
	Undergraduate	38	78	35	82
Statistics Courses Taken	None	17	37	18	39
	One	23	45	15	39
	Multiple	10	15	13	22
Age Group	Under 30 years	40	83	39	84
	30 years and over	10	14	7	16
Has Disability	No	34	76	34	85
	Yes	16	21	12	15

for that measure. For the uncertainty and correctness analyses, only participants who completed all relevant tasks were included to avoid inflating their influence on the averages. Finally, qualitative feedback was roughly analyzed for themes, with attention to differences by software condition to contextualize participant experiences.

3.5 Results

3.5.1 Participant Flow

Participants were randomly assigned to the Experimental or Traditional design group, stratified by gender during the demographics survey. Five participants (two Experimental, three Traditional) selected “Other” or “Prefer not to say” for gender; only three completed the study (two Experimental, one Traditional) and were excluded from quantitative analyses due to insufficient representation.

At the start of the study, there were fewer female than male participants, and this imbalance persisted through retention. In the Experimental group, 46 of 52 females (88.46%) completed the study compared to 86 of 105 males (81.90%). In the Traditional group, 41 of 51 females (80.39%) completed the study versus 91 of 107 males (85.05%). Thus, females had higher retention than males in the Experimental group but lower retention in the Traditional group. The largest retention drop occurred in the selection (tasks 1 and 2) and

input (tasks 3 and 4) stages, most notably among females in the Traditional group (46 to 41 participants, 10.87% drop) compared to males (100 to 92, 8.00% drop). This pattern suggests that certain stages in the Traditional condition may have posed greater challenges for females, contributing to higher attrition. While exact causes of dropout cannot be determined here, the results suggest design-related factors may influence retention differently by gender, warranting further investigation.

3.5.2 Recruitment

Recruitment took place from May 2024 to December 2024, with a higher rate of participants enrolling between September and December. Participants were recruited through a combination of online outreach, departmental emails, disability resource centers, and in-person efforts. A total of 315 non-excluded participants accessed the study, with 293 of them completing at least the first task and thus capable of being included in some analyses. Resulting in nearly equal distribution across the Experimental ($n = 147$) and Traditional ($n = 146$) conditions.

3.5.3 Reliability and Bias Checks

Before testing hypotheses, we conducted a series of checks to evaluate whether our measures were consistent, valid, or biased across conditions. These analyses assessed the stability of self-report ratings, the expected relationship between anxiety and confidence, the reliability of performance grading, task time equivalence, and potential design-related bias in neutral questions.

First, test-retest analyses indicated strong stability of single-item anxiety and confidence ratings across two repeated tasks in each area: selection ($r = .70$ and $r = .67$), input ($r = .82$ and $r = .80$), and output ($r = .82$ and $r = .78$). This does not mean it provides the same information as other instruments, but does demonstrate that with these single item scales, people tend to mark their answers consistently.

Second, anxiety and confidence showed the same negative correlation we expect from prior work on statistics anxiety and self-efficacy [87, 66, 108, 134]. Correlations were moderately negative overall ($r = -.29$) and within each domain (selection $r = -.36$, input $r = -.32$, output $r = -.30$, all $p < .001$). Most sub-scales followed this pattern, with the exception of one weaker sub-scale in the output stage (*The code used to produce output*, $r = -.14$, $p = .021$), suggesting it may have been less central to participants' experience when interpreting the output. This again does not tell us these measures give the same information, but does tell us we observe a similar pattern.

Third, inter-rater reliability checks confirmed that the code evaluation rubric was applied consistently. Across 855 graded criteria in a sample of the dataset, raw agreement between two independent raters was 96% (Cohen's $\kappa = .899$, $SE = .018$), indicating very strong reliability [76]. Once this agreement was established, the remaining responses were graded by a single rater.

Fourth, task completion times were examined to verify that both software designs required similar depth and difficulty. Across all six tasks, participants averaged 27 minutes in total (3 to 5 minutes per task), with

no significant differences between Experimental and Traditional groups, regardless of whether outliers were included ($F(1, 1633) = 0.38, p = .539$) or excluded ($F(1, 1512) = 1.62, p = .204$). These results indicate that tasks took roughly the same amount of time across conditions, ensuring that observed group differences in affective measures were not simply due to easier or at least faster task experiences.

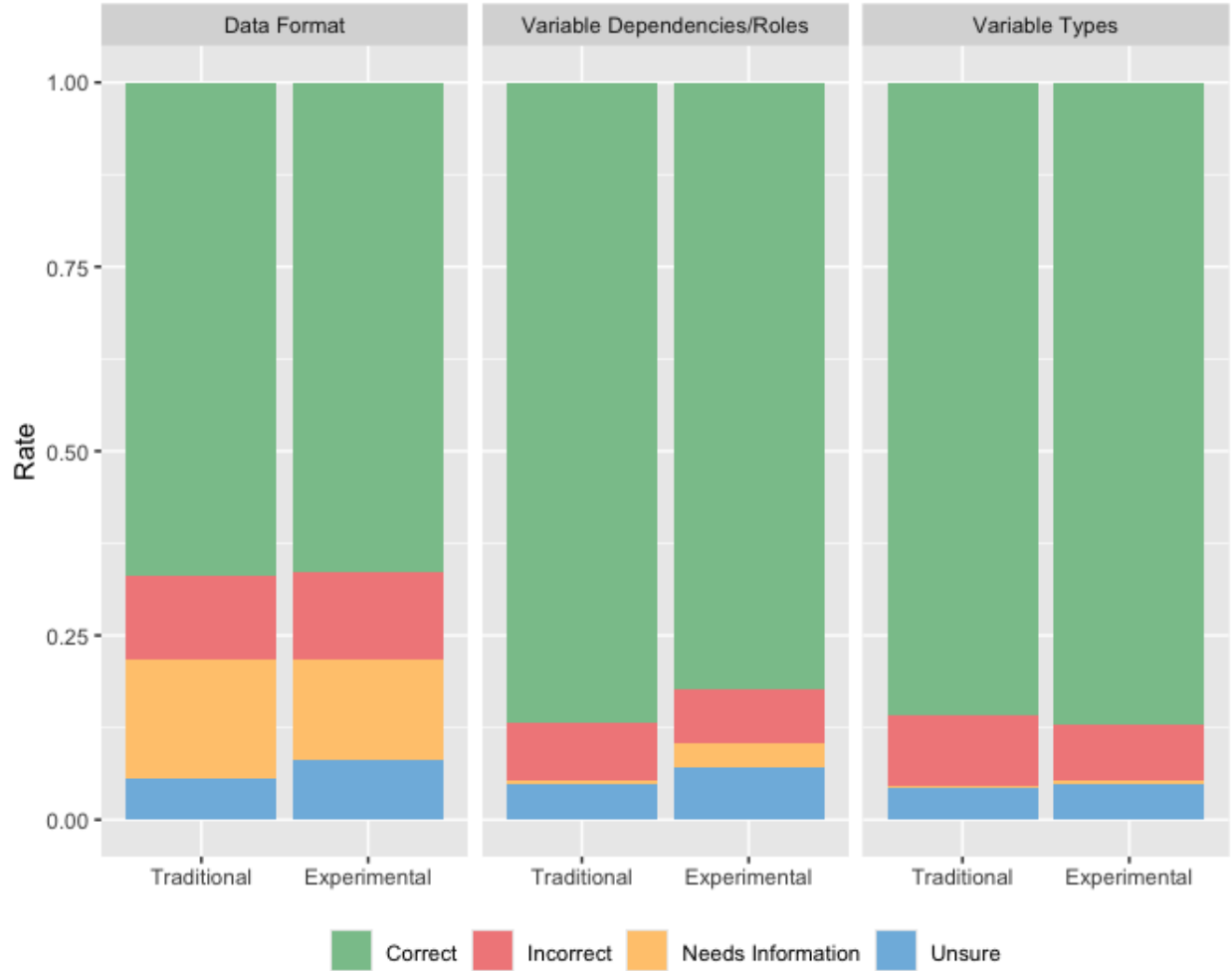


Figure 3.11: Correctness (green) and uncertainty (blue) rates on neutral data-related questions. Performance was similar between Experimental and Traditional groups, indicating that software design did not bias responses on questions unrelated to design.

Finally, we took one more approach to evaluating our single item measures, which was to use neutral questions. Notably, we included ten neutral items in the first two tasks (e.g., identifying data format or variable type) unrelated to software design. Since these were the same, if we observed differences, participants could just be blindly marking. Indeed, no differences were found between Experimental and Traditional groups in uncertainty or correctness ($F(1, 277) = 0.40, p = .525$; $F(1, 277) = 1.00, p = .318$)(Figure 3.11), nor in anxiety or confidence data-related sub-scales ($F(1, 279) = 0.81, p = .369$; $F(1, 279) = 0.34, p = .559$).

This suggests that the observed software design effects are unlikely to reflect differences in participants' comfort with basic data concepts. We note that this check does not account for differences in coding or general software experience, but it provides a focused validation that the core results that cannot be explained by systematic bias in data-related understanding. We still do not think this battery of tests answers all of the questions even we as authors had about using a 1-item scale, but we also thought it was promising compared to the alternatives known in the literature.

Taken together, these analyses provide evidence that the study tasks and measures were reliable, valid, and unbiased across conditions. Full ANOVA tables, correlation tables and rater agreements are provided in Appendix A.2.1 for completeness.

3.5.4 Statistics and Data Analysis

This section reports the results of statistical analyses examining participant performance, uncertainty, anxiety, confidence, and task difficulty across software design conditions.

Analyses are presented in three parts. First, study-wide outcomes are reported, including overall anxiety, confidence, uncertainty, correctness, and perceived difficulty. Second, results are broken down by task area (selection, input, output), with attention to differences in anxiety, confidence, uncertainty, and correctness (correctness in input tasks was graded using the established rubric previously explained). Third, qualitative feedback is summarized to provide context for participants' experiences and to highlight perceptions not captured by quantitative measures.

For clarity, detailed ANCOVA/ANOVA tables and post hoc comparisons that are not found in the main text are provided in Appendix A.2; the main text highlights key effects and patterns relevant to software design and task areas.

3.5.4.1 Study-Wide Analyses

The measures from all three task areas (each of the six tasks) were averaged to examine the overall influence of the effects throughout the statistical analysis process. This averaging mitigates any potential within-subjects effects that may develop as participants progress through the tasks. To ensure unbiased results, only participants who completed all tasks were included in the overall analyses of this section.

Study Completion Rate Completion rate may provide an indicator of participant engagement, though the specific reasons for non-completion cannot be determined. The study was conducted remotely, and participants completed it in varied settings, which may have influenced their ability or motivation to continue. Figure 3.12 shows completion rates at each stage for all participants (red lines) and for participants reporting disabilities (blue lines). Participants with disabilities in the Experimental group demonstrated higher completion rates than their overall group, a pattern observed for both men and women. Conversely, in the Traditional group, participants with disabilities showed lower completion rates. Additionally, women

in the Experimental group had higher completion rates than women in the Traditional group. These trends suggest that study design may interact with participant characteristics, such as disability status or gender, to influence retention.

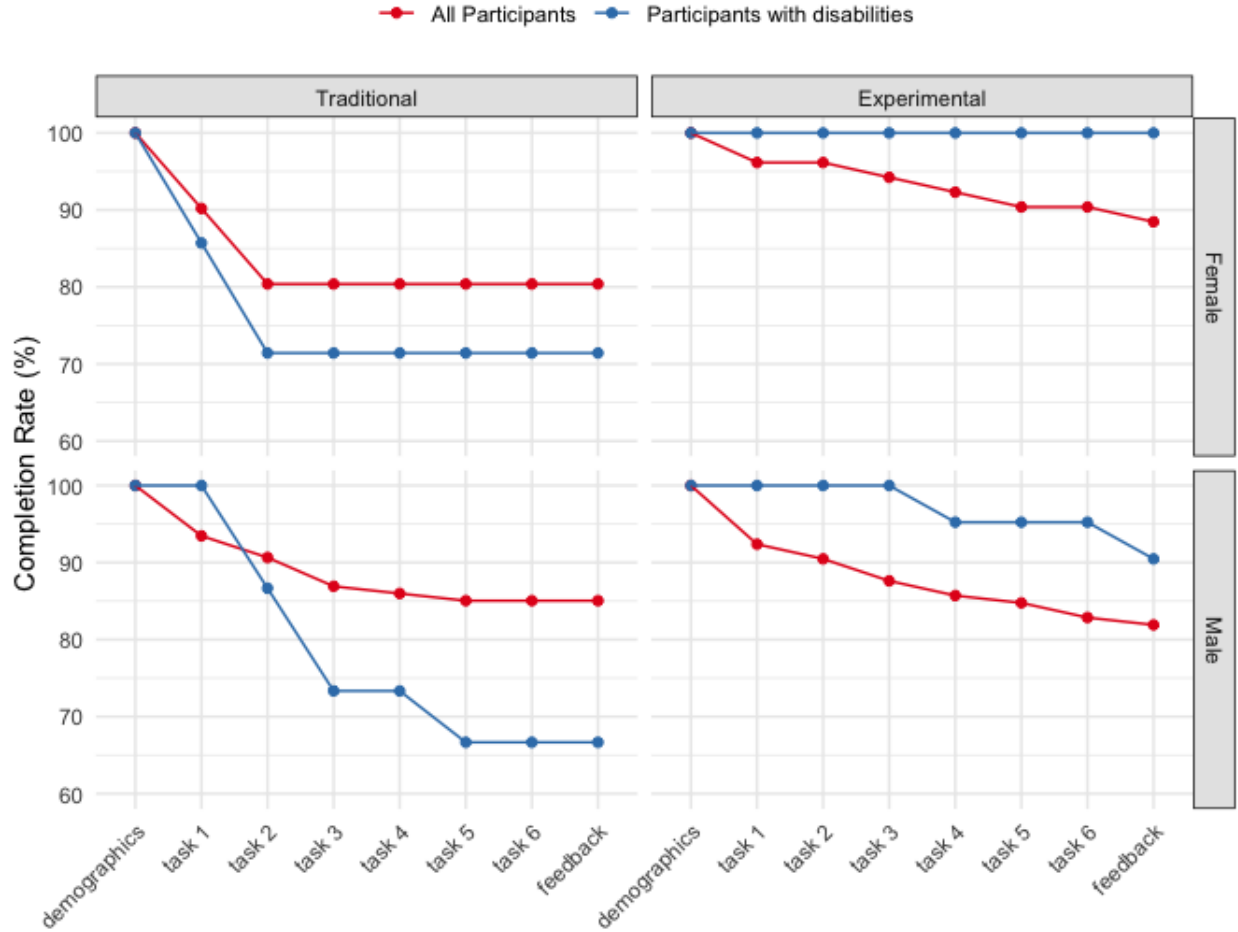


Figure 3.12: Completion rate at each stage of the study

Study-Wide Anxiety Participants' average anxiety across all tasks differed slightly by software design. The Experimental group reported lower anxiety ($M = 2.87, SD = 1.21$) than the Traditional group ($M = 3.10, SD = 1.22$), a small effect ($F(1, 255) = 4.10, p = .051, \eta_p^2 = .010$; Table 3.8). Gender, age and disability status showed marginal or small influences: women ($M = 3.19$ vs. $M = 2.88$), participants with disabilities ($M = 3.25$ vs. $M = 2.91$), and younger participants ($M = 3.03$ vs. $M = 2.71$) tended to report higher anxiety, though effects were weaker ($\eta_p^2 \leq .017$). Other factors, including academic field, experience and language, had negligible impact. **To summarize, both software design and disability status had small effects on study-wide anxiety, while other participant demographics had minimal**

influence. The Experimental group reported 3.29% less anxiety than the Traditional group.

Table 3.8: ANCOVA results for study-wide average anxiety.

ANCOVA Results					
Source	SS	df	F	p-value	Partial η^2
Between Groups					
Design	5.47	1	3.84	0.051	0.010
Gender	4.20	1	2.95	0.087	0.010
Academic Field	0.38	1	0.27	0.607	0.001
Primary Language	0.44	1	0.31	0.580	0.001
Age Group	4.39	1	3.08	0.080	0.011
Has Disability	6.29	1	4.42	0.037*	0.017
Statistics Course Experience	5.98	2	2.10	0.125	0.016
Design $\times$ Gender	1.21	1	0.85	0.356	0.003
Gender $\times$ Academic Field	0.87	1	0.61	0.435	0.002
Within Groups					
Residuals	362.94	255			

Study-Wide Confidence Participants' average confidence across all tasks was influenced by software design. The Experimental group reported higher confidence ($M = 3.35, SD = 1.20$) than the Traditional group ($M = 3.06, SD = 1.20$), a statistically significant effect ($F(1, 255) = 7.50, p = .007, \eta_p^2 = .029$; Table 3.9). Participant experience in statistics courses, gender and age also had small but notable effects: those with more course experience, older participants and men generally reported higher confidence. Post hoc comparisons showed that those who have taken multiple statistics course reported the most confidence ($M = 3.62, SD = 1.38$). Notably, there was not a significant difference in confidence between those who took only one course and those who never took one ($p = .487$). Gender effects varied by academic field, with women in formal sciences reporting higher confidence ($M = 3.39, SD = 1.24$) than women in other fields ($M = 2.71, SD = 1.18$) ($p = .024$), while men's confidence was consistent across fields ($p = .980$). In formal sciences, men and women both reported similar confidence levels ($p = .993$). **To summarize, software design increased study-wide confidence, and prior statistics experience, age and gender had additional influences, although gender differences varied by academic field. The Experimental group reported 4.14% higher confidence than the Traditional group.**

Study-Wide Uncertainty Participants' overall uncertainty (i.e., the rate of selecting "I am not sure") was strongly influenced by software design. Those using the Traditional design were more uncertain ($M =$

Table 3.9: ANCOVA results for study-wide average confidence.

ANCOVA Results					
Source	SS	df	F	p-value	Partial η^2
Between Groups					
Design	9.55	1	7.50	0.007**	0.029
Gender	8.23	1	6.46	0.012*	0.025
Academic Field	2.03	1	1.59	0.208	0.006
Primary Language	1.27	1	1.00	0.319	0.004
Age Group	6.36	1	4.998	0.026*	0.019
Has Disability	2.22	1	1.74	0.188	0.007
Statistics Course Experience	17.33	2	6.80	0.001**	0.051
Design $\times$ Gender	3.39	1	2.66	0.104	0.010
Gender $\times$ Academic Field	8.60	1	6.75	0.010**	0.026
Within Groups					
Residuals	324.70	255			

41.2, $SD = 23.9$) than those using the Experimental design ($M = 25.0$, $SD = 23.3$), a large and statistically significant effect ($F(1, 253) = 34.76$, $p < .001$, $\eta_p^2 = .120$; Table 3.10). Other notable predictors included experience in statistics courses and participants' confidence: taking at least one statistics course and having higher confidence was associated with lower uncertainty, accounting for 9.6% and 17.4% of the variance, respectively. Age had a minor influence, with younger participants reporting higher uncertainty ($M = 34.6$ vs. $M = 24.4$), while gender, disability, and academic field had negligible effects. Furthermore, confidence shared a strong negative correlation with uncertainty ($r = -.49$, $p < .001$) and anxiety a weak positive one ($r = .19$, $p = .002$). **In other words, software design has a large impact on participants' engagement and uncertainty when attempting statistical tasks, with prior course experience and confidence contributing similar influences. The Traditional group was 16.20% more uncertain, though they had more statistical experience on average than the Experimental group.**

Study-Wide Correctness Correctness was an objective measure and here too we saw differences. Participants' accuracy, measured as the proportion of correct responses for attempted questions, was strongly influenced by software design. Those using the Experimental design outperformed the Traditional group ($M = 67.1$, $SD = 16.8$ vs. $M = 56.0$, $SD = 14.7$), a large and statistically significant effect ($F(1, 253) = 35.52$, $p < .001$, $\eta_p^2 = .120$; Table 3.11). Other factors also had modest effects on correctness. Prior statistics course experience and reported confidence were associated with higher accuracy, with those having taken at least one course performing better ($M = 63.9$ to 67.4) than those with none ($M = 55.5$). Academic field

Table 3.10: ANCOVA results for study-wide average uncertainty.

ANCOVA Results					
Source	SS	df	F	p-value	Partial η^2
Between Groups					
Design	13505.00	1	34.76	< .001***	0.120
Gender	33.00	1	0.08	0.771	0.000
Academic Field	1035.00	1	2.66	0.104	0.010
Primary Language	142.00	1	0.37	0.546	0.001
Age Group	1337.00	1	3.44	0.065	0.013
Has Disability	337.00	1	0.87	0.353	0.003
Statistics Course Experience	11309.00	2	14.55	< .001***	0.096
Average Confidence	20259.00	1	52.15	< .001***	0.174
Average Anxiety	15.00	1	0.04	0.843	0.000
Design $\times$ Gender	537.00	1	1.38	0.241	0.005
Gender $\times$ Academic Field	28.00	1	0.07	0.790	0.000
Within Groups					
Residuals	98290.00	253			

and age group showed small but significant influences, with participants in formal sciences ($M = 63.3$ vs. $M = 59.7$) and those 30 or older ($M = 68.5$ vs. $M = 60.3$) slightly outperforming their peers. Anxiety, gender, and the interaction between gender and academic field did not meaningfully affect performance. **In simple terms, software design had the largest impact on participants' correctness (the Experimental group was 11.10% more accurate even after accounting for uncertainty), while prior course experience, confidence, age, and academic field contributed smaller, supportive effects.**

Perceived Difficulty Participants rated the difficulty of tasks across the three areas (selection, input, output) on a 5-point Likert scale, from 'very easy' (1) to 'very difficult' (5). Task area strongly influenced perceived difficulty, accounting for 34.0% of the variance ($F(2, 783) = 203.16, p < .001, \eta_p^2 = .340$; Table 3.12). Output tasks were perceived as the most difficult ($M = 4.10, SD = 0.94$), followed by input ($M = 3.52, SD = 1.12$) and selection ($M = 2.38, SD = 1.00$) tasks.

Software design also had a significant effect on difficulty ratings ($F(1, 783) = 22.60, p < .001, \eta_p^2 = .030$), and this effect interacted with task area ($F(2, 783) = 10.47, p < .001, \eta_p^2 = .030$). Post hoc comparisons (Table 3.13) revealed that the Experimental group rated input ($M = 3.23, SD = 1.15$) and output ($M = 3.82, SD = 1.00$) tasks as significantly less difficult than the Traditional group ($M = 3.80, SD = 1.02$ and $M = 4.38, SD = 0.79$, respectively), representing roughly an 11% reduction in perceived difficulty. Selection

Table 3.11: ANCOVA results for study-wide average correctness after removing unsure responses.

ANCOVA Results					
Source	SS	df	F	p-value	Partial η^2
Between Groups					
Design	6693.00	1	35.52	< .001***	0.120
Gender	330.00	1	1.75	0.187	0.007
Academic Field	1440.00	1	7.64	0.006**	0.030
Primary Language	610.00	1	3.24	0.073	0.010
Age Group	1481.00	1	7.86	0.005**	0.030
Has Disability	476.00	1	2.53	0.113	0.010
Statistics Course Experience	3231.00	2	8.57	< .001***	0.060
Average Confidence	5058.00	1	26.84	< .001***	0.100
Average Anxiety	496.00	1	2.63	0.106	0.010
Design $\times$ Gender	8.00	1	0.04	0.836	0.000
Gender $\times$ Academic Field	88.00	1	0.47	0.496	0.002
Within Groups					
Residuals	47674.00	253			

tasks did not differ between groups. Gender had no meaningful effect on perceived difficulty. This interaction is illustrated in Figure 3.13. **In summary, task area was the primary driver of difficulty perceptions, but software design meaningfully reduced perceived difficulty in the more challenging input and output tasks, highlighting the benefit of the Experimental interface in easing engagement with these areas. There was no observed difference in how difficult men and women perceived these tasks to be.**

3.5.4.2 Selection Area Analyses

As a reminder, the selection stage of the study was designed to assess participants' ability to choose appropriate statistical tests based on given scenarios. They selected from a list of tests for several scenarios, evaluating their decision-making with different data types and questions. The analysis explored how participants felt while completing these tasks and how statistical knowledge, demographics, anxiety, and confidence influenced accuracy. We explored differences between the Traditional and Experimental groups, especially regarding the impact of naming conventions on performance and perceptions. Full result tables can be found in Appendix A.2.3.

Table 3.12: ANOVA results for perceived task difficulty.

ANOVA Results					
Source	SS	df	F	p-value	Partial η^2
Between Groups					
Design	22.59	1	22.60	< .001***	0.030
Gender	0.01	1	0.01	0.919	0.000
Task Area	406.11	2	203.16	< .001***	0.340
Design $\times$ Gender	0.82	1	0.82	0.365	0.001
Gender $\times$ Task Area	0.52	2	0.26	0.770	0.001
Design $\times$ Task Area	20.92	2	10.47	< .001***	0.030
Design $\times$ Gender $\times$ Task Area	2.25	2	1.13	0.325	0.003
Within Groups					
Residuals	782.59	783			

Table 3.13: Post-hoc comparisons for design and task area on perceived task difficulty

Tukey-Kramer Results	
Design $\times$ Task Area	p-value
Traditional Selection - Experimental Selection	0.9995
Traditional Input - Experimental Input	0.0002 ***
Traditional Output - Experimental Output	0.0004 ***

Selection Anxiety Selection area anxiety reflects participants' average anxiety across the two selection tasks. Software design had no meaningful effect on anxiety in this area ($F(1, 282) = 0.002, p = .967$), indicating that the Experimental and Traditional interfaces elicited similar levels of anxiety during task selection. Other factors showed small but significant effects: participants with disabilities reported higher anxiety ($F(1, 282) = 10.46, p = .001, \eta_p^2 = .036$; $M = 2.79, SD = 1.11$) than those without ($M = 2.31, SD = 1.00$), and gender had a marginal impact, with women slightly more anxious than men ($F(1, 282) = 3.495, p = .063, \eta_p^2 = .012$). A secondary analysis focused on sub-scale ratings for test selection and naming conventions confirmed that design remained non-significant ($F(1, 282) = 1.07, p = .301$), whereas having a disability ($M = 3.19$ vs. $M = 2.58$) and prior statistics coursework were the primary predictors of anxiety in this context. The gender effect diminished, indicating that differences in anxiety between men and women may be more prominent in data- and information-related aspects. **To summarize, the type of software had minimal impact on anxiety for selection tasks even when accounting for neutral aspects; participants' disability status and possibly their statistical experience played a bigger role in shaping their responses.**

Selection Confidence Selection area confidence reflects participants' average confidence across the two selection tasks. The overall analysis showed a significant interaction between software design and gender ($F(1, 282) = 11.43, p < .001, \eta_p^2 = .039$): men in the Traditional group reported the highest confidence ($M = 4.38, SD = 1.16$), while women in the Traditional group reported the lowest ($M = 3.34, SD = 1.17$). In contrast, confidence levels were similar for men and women in the Experimental group ($p = .999$), suggesting that the Experimental design mitigated gender differences in confidence ($M = 3.87, SD = 1.26$ and $M = 3.83, SD = 1.26$, respectively).

Disability status and prior statistics coursework also influenced overall selection confidence. Participants with disabilities reported slightly lower confidence ($F(1, 282) = 3.85, p = .051, \eta_p^2 = .013; M = 3.67$ vs. $M = 4.03$). Confidence increased with more prior statistics coursework ($F(2, 282) = 6.32, p = .002, \eta_p^2 = .043$), driven primarily by participants who had taken multiple courses; participants with only one course did not differ significantly from those with no coursework ($p = .269$), suggesting that a single course may not provide enough experience to noticeably boost confidence.

When focusing specifically on confidence related to the selection of the correct statistical test and naming conventions, software design emerged as a small but significant predictor ($F(1, 282) = 6.78, p = .010, \eta_p^2 = .023$): participants in the Experimental group reported higher confidence ($M = 3.32, SD = 1.38$) than those in the Traditional group ($M = 2.99, SD = 1.46$) across both men and women. This indicates that while design had minimal impact on anxiety, the nomenclature improvements in the Experimental interface positively supported participants' confidence in the selection stage. Age group also explained some variance in nomenclature sub-scale confidence, with older participants reporting higher confidence ($F(1, 282) = 11.00, p = .001, \eta_p^2 = .038; M = 3.83$ vs. $M = 3.03$) and course experience remained a predictor. **In other words, the Experimental design helped boost selection confidence by about 5%, especially for women, reducing gender gaps, while prior experience, age, and disability status continued to play a role.**

Selection Uncertainty Selection area uncertainty reflects participants' rate of "I am not sure" responses across the two selection tasks. Software design had a strong effect in this area ($F(1, 270) = 26.19, p < .001, \eta_p^2 = .090$), with participants in the Experimental group ($M = 9.1, SD = 14.8$) showing half the uncertainty of those in the Traditional group ($M = 17.5, SD = 15.2$). Gender had a smaller but significant effect ($F(1, 270) = 6.76, p = .010, \eta_p^2 = .020$), with men reporting slightly higher uncertainty ($M = 14.1, SD = 15.6$) than women ($M = 11.3, SD = 15.4$).

Prior statistics coursework also influenced uncertainty ($F(2, 270) = 13.04, p < .001, \eta_p^2 = .090$), with participants who had taken no courses showing the highest uncertainty ($M = 19.4, SD = 19.0$) and those with one or multiple courses reporting lower, similar levels of uncertainty ($M = 10.7$ and $M = 7.4$, respectively; $p = .325$). Confidence was negatively associated with uncertainty ($F(1, 270) = 16.54, p < .001, \eta_p^2 = .060; r = -.27, p < .001$), whereas anxiety was not a significant predictor ($F(1, 270) = 0.09, p = .764$).

Uncertainty was virtually non-existent in the neutral data-related questions in either group (shown previously in Figure 3.11). So, a secondary look at individual test-selection questions showed that design effects were particularly pronounced: participants in the Experimental group were more decisive, with 31.7%, 37.9%, and 44.90% lower uncertainty when selecting the Pearson correlation, Pearson chi-squared, and ANOVA tests, respectively (shown in Figure 3.14). **To summarize, software design was the primary factor reducing uncertainty in selection tasks, Experimental participants were about 38.2% more willing to commit to answers on test-selection questions. While demographic factors such as gender and prior statistics experience had smaller effects, design improvements appear to be the dominant influence on decisiveness in this area.**

Selection Correctness Selection area correctness reflects how accurately participants chose the best answer in the selection tasks, calculated only on attempted responses (excluding “I am not sure”), to give a clearer measure of actual performance. When considering the average across all 13 selection area questions, correctness did not vary greatly across design or demographics. Participants in the Traditional group achieved slightly higher correctness ($M = 82.1\%$, $SD = 16.4$) compared to the Experimental group ($M = 78.0\%$, $SD = 14.6$), but this difference was marginal ($F(1, 268) = 3.36, p = .068, \eta_p^2 = .010$). Formal science participants performed better ($M = 82.0\%$) than those in other areas (77.6%) ($F(1, 268) = 4.93, p = .027, \eta_p^2 = .020$), while gender did not yield differences. Confidence was a predictor and positively linked to correctness ($r = .21, p < .001$), while anxiety showed a small negative relationship ($r = -.17, p = .005$), suggesting that individual factors influenced performance but did not dramatically alter overall correctness.

More notable differences appeared at the level of specific test-selection questions, where design affected both whether participants engaged and how well they performed once they engaged differently, so we observed correctness in two ways per test question: overall accuracy (all participants) and conditional accuracy (only those who attempted an answer). Naming conventions had a clear impact, sometimes in surprising ways. For example, with the chi-squared test, participants in the Experimental group were more likely to attempt an answer (CompareCounts: 73.5% vs. chisq.test: 35.6%), but they were less successful when they did (12.0% vs. 32.6% accuracy among those who tried). This meant that overall accuracy was similarly low (8.8% vs. 11.6%), but of those that tried, the Traditional group was more accurate (see Figure 3.14). A similar trade-off appeared with the correlation test: CorrelateGroups nearly doubled engagement compared to cor.test (71.4% vs. 39.7%), yet the proportion of correct answers among those who engaged was lower (41.0% vs. 60.3%), producing only a small gain in overall accuracy (29.3% vs. 24.0%). In contrast, clearer, more intuitive names yielded consistent improvements: CompareMeans drove both higher engagement (86.2% vs. 41.3%) and higher overall performance (55.5% vs. 21.0%). **In summary, some naming choices increased willingness to engage without improving correctness, while others improved both engagement and correctness. The results suggest that test names which clearly convey purpose can meaningfully improve both engagement and success, whereas names based only**

on mathematical terminology may obscure meaning and provide little benefit over traditional labels.

3.5.4.3 Input Area Analyses

As a brief reminder, the input area assessed participants' experiences with implementing statistical tests through code, focusing on tasks involving writing, modifying, or interpreting statistical code. The purpose in including such a task is at least because scientists often use code to ensure others can replicate their findings. The section similarly measured anxiety, confidence, and accuracy, while examining variation across design groups, prior statistics experience, and demographics. Full result tables can be found in Appendix A.2.4.

Input Anxiety Input area anxiety reflects participants' average anxiety across the two coding tasks. Software design showed a marginal effect in this area when analyzing the main anxiety scale ($F(1, 260) = 3.53, p = .061, \eta_p^2 = .010$) but proved to be a stronger predictor when considering only code-related aspect sub-scales ($F(1, 260) = 8.12, p = .005, \eta_p^2 = .030$), with participants in the Traditional group reporting higher anxiety in code aspects ($M = 3.30, SD = 1.66$) than those in the Experimental group ($M = 2.80, SD = 1.52$). Age and gender differences also appeared in the overall ratings: younger participants experiencing more anxiety than older ones ($F(1, 260) = 5.14, p = .024, \eta_p^2 = .020$; $M = 3.40$ vs. $M = 2.83$) and women reporting more anxiety than men ($F(1, 260) = 4.86, p = .028, \eta_p^2 = .020$; $M = 3.65$ vs. $M = 3.14$), but these effects diminished in the code-related aspect ratings. The gender differences in this stage are consistent with prior findings on gender disparities in programming-related or computer anxiety. However, the larger effect in the overall rating analysis suggests that women may be affected by an external variable unrelated to design or not fully captured in the sub-scales.

The secondary analysis of code-related factors (e.g., data format, examples, explanation, and complexity) revealed that design affected men and women differently, with post hoc analysis showing that the Experimental design significantly lowering anxiety for men ($p = .002$) in particular. Differences by academic field remained, with participants in the formal sciences reporting lower anxiety ($M = 2.86, SD = 1.51$) than those in other areas ($M = 3.25, SD = 1.70$). **Put simply, the Experimental design clearly lowered input anxiety, especially for code-related aspects (7.14% less anxiety than the Traditional group), while age, gender, and academic field played roles as well. Because only two simple and common statistical methods were considered and precise code examples were given in this study, these effects may underestimate the challenges users would face in more complex analyses, where design differences could have an even greater impact.**

Input Confidence Input area confidence reflects participants' average confidence across the two coding tasks. Software design had a clear effect: participants in the Experimental group reported higher confidence ($M = 3.09, SD = 1.55$) than those in the Traditional group ($M = 2.66, SD = 1.60$), indicating that the

Experimental interface led to a higher reporting of confidence ($F(1, 260) = 8.25, p = .004, \eta_p^2 = .030$).

Demographic factors also played a role. Men reported more confidence ($M = 3.05, SD = 1.56$) than women ($M = 2.54, SD = 1.59$), and older participants felt more confident ($M = 3.32, SD = 1.52$) than younger ones ($M = 2.80, SD = 1.59$). Academic background and statistics experience mattered as well: participants in the formal sciences ($M = 3.09$ vs. $M = 2.64$) and those with multiple prior statistics courses reported the highest confidence ($M = 3.30, SD = 1.79$), while those with no coursework reported the lowest ($M = 2.57, SD = 1.57$). A gender-academic-field interaction with post hoc analysis revealed that women in fields other than formal sciences felt significantly less confident than all other participants (all with $p < .006$), and that gender gaps in confidence were not found within the formal sciences ($p = .950$). A secondary analysis of code-related confidence (e.g., format, examples, naming, and complexity) conveyed similar results; that design, prior statistics coursework, academic field, and gender, significantly shaped participants' coding confidence. **In other words, the Experimental design measurably boosted participants' confidence in statistical coding tasks (6.29% more confident than the Traditional group), while demographic and experiential factors still added meaningful contributions. Women in non-formal science fields experienced the lowest coding confidence.**

Input Uncertainty Uncertainty in the Input area was measured as the proportion of coding tasks where participants responded "I am not sure." Because only two coding questions were included, scores clustered at 0%, 50%, or 100%. Even with this tri-chotomized measure, a clear design effect emerged: the Experimental group showed lower uncertainty overall (33.3%) compared to the Traditional group (49.2%). Women were more uncertain (51.1% vs. 36.3%), but particularly women in the Traditional group reported the highest uncertainty (53.7%), whereas men in the Experimental group reported the lowest (25.0%). Figure 3.15 illustrates these group-level patterns, showing that uncertainty rates were stable across tasks regardless of design.

Participants with more statistical coursework were less uncertain: those with multiple courses had the lowest uncertainty (27.3%), followed by those with one course (35.3%), while participants with no prior coursework had the highest uncertainty (56.2%). Younger participants (43.2% vs. 29.5%) and participants in non-mathematical fields were more uncertain (49.2% vs. 33.7%). A closer look at task type showed that uncertainty was consistently higher in the Traditional group when implementing both one-way and repeated-measures ANOVAs, with the Experimental group showing 14.9 and 16.9% lower uncertainty in each case. Spearman correlational analyses aligned with these findings: lower confidence was strongly linked to higher uncertainty ($r_s = -.45, p < .001$), and greater anxiety was moderately linked to uncertainty ($r_s = .26, p < .001$). **In simple terms, the Experimental design reduced participants' uncertainty by 15.9% when coding and this advantage held across both tasks. Demographics and prior experience shaped uncertainty as well, but a clear pattern was that the redesigned interface helped participants feel more certain in their coding decisions.**

Input Correctness Input area correctness reflects the percentage of correct coding criteria among attempted answers (excluding “I am not sure” responses). Software design showed a clear effect: participants in the Experimental group achieved higher correctness ($M = 81.9\%$, $SD = 20.8$) than those in the Traditional group ($M = 67.4\%$, $SD = 25.5$), ($F(1, 164) = 12.02, p < .001, \eta_p^2 = .070$). Academic field also mattered, with participants in the formal sciences performing more accurately ($M = 77.9\%$, $SD = 23.0$) than those from other fields ($M = 72.3\%$, $SD = 25.2$), ($F(1, 164) = 4.77, p = .030, \eta_p^2 = .030$); considering that many of the formal science participants majored or were majoring in Computer Science and were likely somewhat familiar with code, this is not surprising. Confidence was a significant predictor ($F(1, 164) = 6.08, p = .015, \eta_p^2 = .040$), moderately correlated with correctness ($r = .26, p < .001$) while anxiety had only a marginal relationship ($r = -0.14, p = .056$). Other demographics and gender showed no effects.

A closer examination of grading criteria highlights how design influenced specific aspects of correctness. Both groups performed well on basic setup tasks such as creating a data frame and specifying filenames (>90% accuracy). However, the Experimental group consistently outperformed the Traditional group on more complex, integrative steps. For instance, they were more accurate in specifying the main test name and parameters (84% vs. 40%), avoiding syntax errors (68% vs. 42%), and ensuring no missing elements (85% vs. 55%). These results suggest that the Experimental design supported participants in organizing and executing full statistical tests. However, the Traditional group showed advantages on subtler, syntax-specific elements. They more often succeeded in naming repeated measures factors correctly (97% vs. 75%) and in formatting repeated factor lists with the correct quotation syntax (96% vs. 74%). Thus, while the Experimental design improved higher-order logic and integration, it sometimes introduced confusion around structural conventions that may have differed from alternative syntax, especially regarding data with repeated measures. Interestingly, this evidence suggests a plausible validation of the Mayrhauser & Vans theory of program comprehension, although re-imagined in a very different context [140].

Importantly, across both designs, participants struggled to produce fully runnable code: only 29-57% of Task 3 responses and 41-49% of Task 4 responses were correct end-to-end. This likely reflects the challenge of programmatic precision in a non-runnable environment, where small errors could not be debugged, a limitation of the present study. To mitigate this, each task included an explicit example that could be copied, requiring only variable substitution, to ensure participants had a working template despite the lack of iterative testing. **In summary, the Experimental design boosted correctness by 14.50% overall, but especially for higher-order and integrative coding steps, while the Traditional design held advantages in narrow syntax conventions. Academic field and confidence further influenced accuracy, but gender and anxiety did not. In more complex tasks, the correctness gap would likely widen, whereas in live coding settings an additional measure (e.g., time to reach the correct answer) would be needed to capture the full impact of design and syntax.**

3.5.4.4 Output Area Analyses

As a reminder, the output area of the study focuses on participants' ability to interpret statistical output, specifically from t-tests and ANOVAs. This includes understanding the significance of results, effect sizes, and statistical labels (such as p-values, assumption tests, and post hoc analyses). In R, output typically appears as compact tables with labels or numerical summaries such as “ $t = -2.31$, $df = 45$, $p\text{-value} = 0.526$ ”, requiring users to recall conventions and thresholds. By contrast, the experimental system generated short narrative statements and APA-style reporting, for example: “A comparison was conducted between two groups. Results show that groupA it is significantly higher than groupB, $t(45) = -2.31$, $p = .026$ ”. We chose these statistical methods largely because they are commonly reported in computing education, including introductory statistics for CS students. The aim is to assess how well participants can translate raw statistical output into meaningful conclusions, as this is crucial for effective data analysis in research and applied settings. Full ANOVA tables for the analyses can be found in Appendix A.2.5.

Output Anxiety Output area anxiety reflects participants' reported stress or discomfort when interpreting statistical results. Software design significantly influenced anxiety: participants using the Experimental output format, which simplified results with descriptive sentences and an approximately eighth-grade reading level, reported lower anxiety ($M = 3.06$, $SD = 1.47$) than those using the Traditional format ($M = 3.49$, $SD = 1.56$; $F(1, 257) = 6.36$, $p = .012$, $\eta_p^2 = .020$). Other demographic factors were mostly non-significant, though participants reporting a disability tended to experience slightly higher anxiety ($M = 3.51$ vs. $M = 3.21$), hinting at potential accessibility challenges with how statistical output is presented. Gender, academic field, age, primary language, and prior statistics coursework did not significantly influence output anxiety. A sub-scale analysis of specific output aspects showed no significant interactions with design, suggesting that while the Experimental format reduced anxiety overall, the exact feature driving this effect could not be isolated. **In other words, simplifying statistical output in this way reduced anxiety across the board by 6.14%, benefiting participants who might find interpreting traditional outputs intimidating, such as those with disabilities. The effect was consistent across different parts of the output, though the study could not pinpoint which specific element was most helpful.**

Output Confidence Output area confidence reflects participants' self-assessed ability to interpret statistical results across both output tasks. Software design had a clear effect: participants using the Experimental output format reported higher confidence ($M = 3.00$, $SD = 1.42$) than those using the Traditional format ($M = 2.47$, $SD = 1.46$; $F(1, 257) = 12.81$, $p < .001$, $\eta_p^2 = .050$). Prior statistics experience also influenced confidence. Participants who had taken multiple statistics courses reported the highest confidence ($M = 3.23$, $SD = 1.62$), whereas those who had taken one or no courses showed similar confidence levels ($M = 2.46 - 2.74$; $p = .707$). Other factors, including gender, academic field, and interactions between them, did not significantly impact confidence, though a marginal trend suggested that men and women may differ

in confidence depending on their discipline.

Analysis across output sub-scales (e.g., table layout, labels, explanation text) did not reveal any significant interactions with design, indicating that the overall format and presentation style of the output, rather than individual elements, drove the observed confidence gains. **In summary, the Experimental output format increased participants' confidence in interpreting results by 7.57%, while experience with statistics also played a role. The effect appears to stem from the overall clarity and readability of the output rather than any single aspect that was measured.**

Output Uncertainty Output area uncertainty reflects participants' reported likelihood of being unsure about their interpretation of results. Software design made a large impact: participants using the Experimental output format reported lower uncertainty ($M = 25.7, SD = 23.7$) than those using the Traditional format ($M = 45.6, SD = 27.5$; $F(1, 253) = 32.80, p < .001, \eta_p^2 = .110$), suggesting that a clearer, simplified output layout helps users feel more certain when interpreting statistical results (Figure 3.16).

Confidence was the strongest predictor of uncertainty. Higher confidence was associated with lower uncertainty ($F(1, 253) = 71.46, p < .001, \eta_p^2 = .220; r = -0.53$), indicating that participants who felt more capable navigating output were less likely to report being unsure. Anxiety was weakly related to uncertainty ($r = 0.16, p = .010$), but did not significantly predict it in the full model, suggesting a modest connection between anxious feelings and uncertainty.

Prior statistics experience also influenced uncertainty: participants who took at least one course reported lower uncertainty ($M = 32.1 - 40.3$) compared to those with no coursework ($M = 55.3$; $F(2, 253) = 7.49, p < .001, \eta_p^2 = 0.060$). Gender and interactions between design and gender were not significant, indicating that uncertainty was primarily shaped by the output format and participants' confidence rather than demographic factors. **To summarize, the Experimental output design reduced uncertainty by 19.90% on average. Participants with higher confidence or more statistics experience also engaged more, while gender and other demographics had little effect. Output intended to be clear and readable appears to help users feel more certain about their interpretations and appears to have helped keep them engaged with the task.**

Output Correctness Output area correctness reflects participants' accuracy in selecting the best answers across the two tasks (9 questions total), considering only attempted responses and excluding any "I am not sure" answers. By this we mean are measuring overall and not conflating uncertainty with errors. Software design significantly influenced correctness, with participants in the Experimental group achieving higher accuracy ($M = 71.0, SD = 23.1$) than those in the Traditional group ($M = 61.9, SD = 25.8$; $F(1, 216) = 7.14, p = .008, \eta_p^2 = .030$), suggesting that the Experimental output format supports more accurate interpretation.

Age was a notable predictor: older participants performed better ($M = 77.7, SD = 21.3$) than younger participants ($M = 64.4, SD = 24.9$; $F(1, 216) = 12.09, p < .001, \eta_p^2 = .050$), possibly reflecting greater

experience or exposure to statistical concepts. Prior statistics coursework showed a marginal trend, with participants who had taken multiple courses performing slightly better than those with one or no courses, although post hoc comparisons were not significant. Confidence was moderately correlated with correctness ($r = .25, p < .001$), indicating that participants who felt more confident also tended to choose more correct answers. Anxiety showed a small negative correlation ($r = -0.13, p = .042$) but did not significantly predict correctness in the model. As with previous analyses, gender did not significantly affect correctness. **In simple terms, the Experimental output design increased not only engagement, but also correctness when engaged by 9.1% on average. Age and confidence contributed meaningfully to performance, while gender and anxiety had little effect, and prior coursework only showed a weak trend. These results suggest that simplifying output and presenting results in a clear, readable format improves participants' ability to correctly interpret statistical information.**

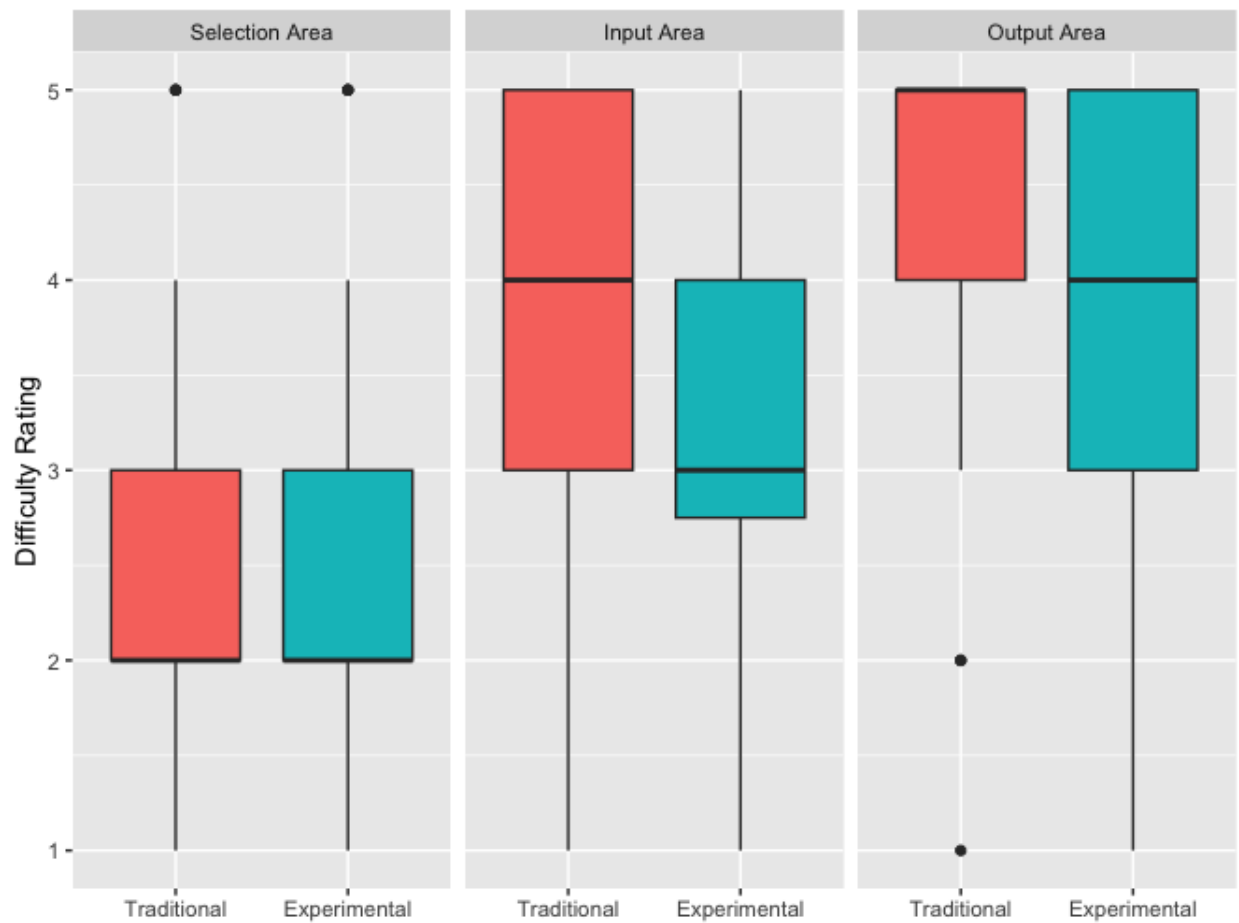


Figure 3.13: Perceived difficulty for each task area. Participants rated how challenging the two tasks in each area were on a 5-point scale (1 = very easy, 5 = very difficult). The figure highlights that participants using the Experimental design rated input and output tasks as less difficult than those using the Traditional design, while selection tasks showed no difference.

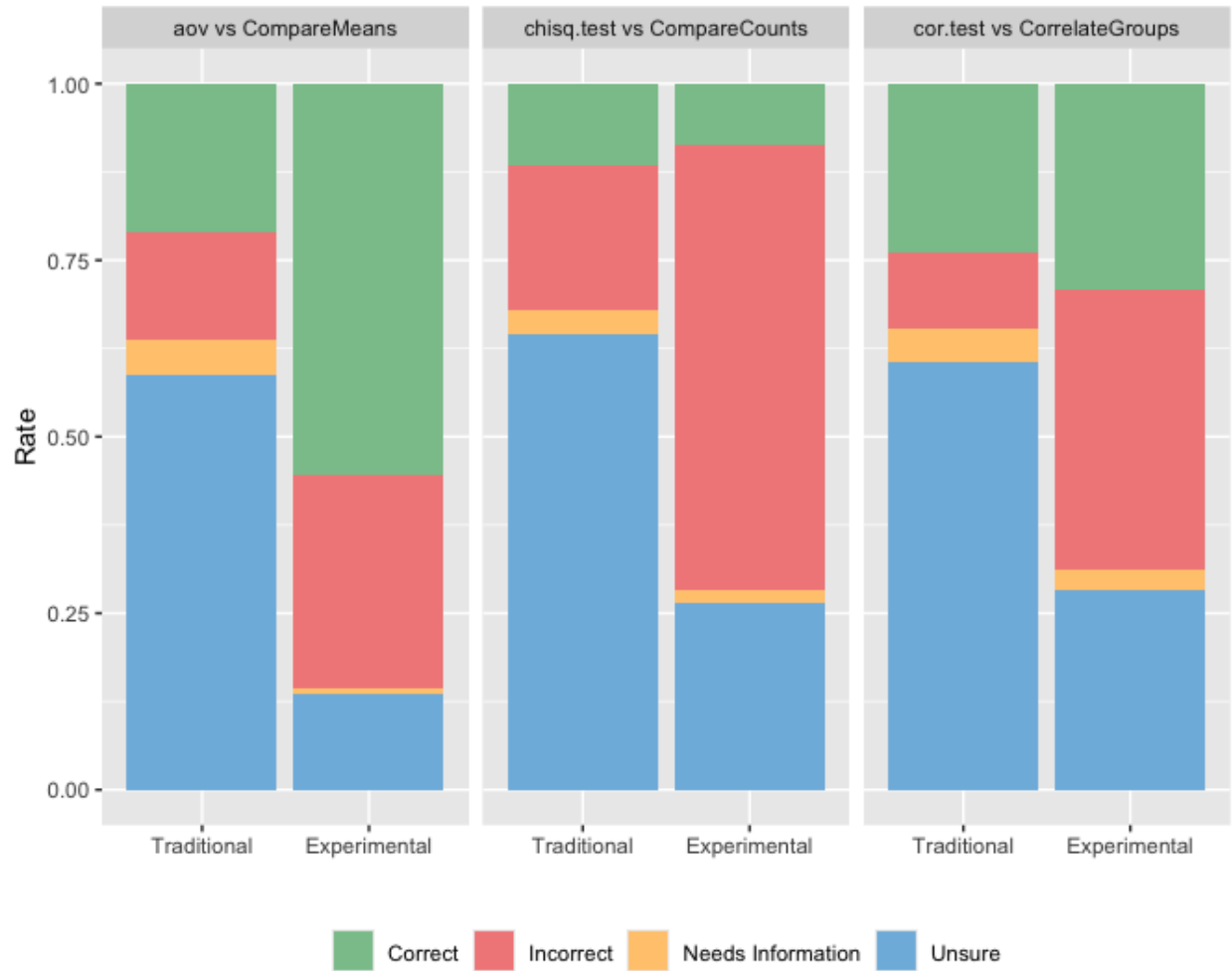


Figure 3.14: Correctness (green) and uncertainty (blue) rates on design-dependent questions. Uncertainty was much higher in the Traditional group and correctness was similar between Experimental and Traditional groups in two of the three questions, indicating that the naming convention can promote engagement but needs more thoughtful and specific consideration to also improve accuracy.

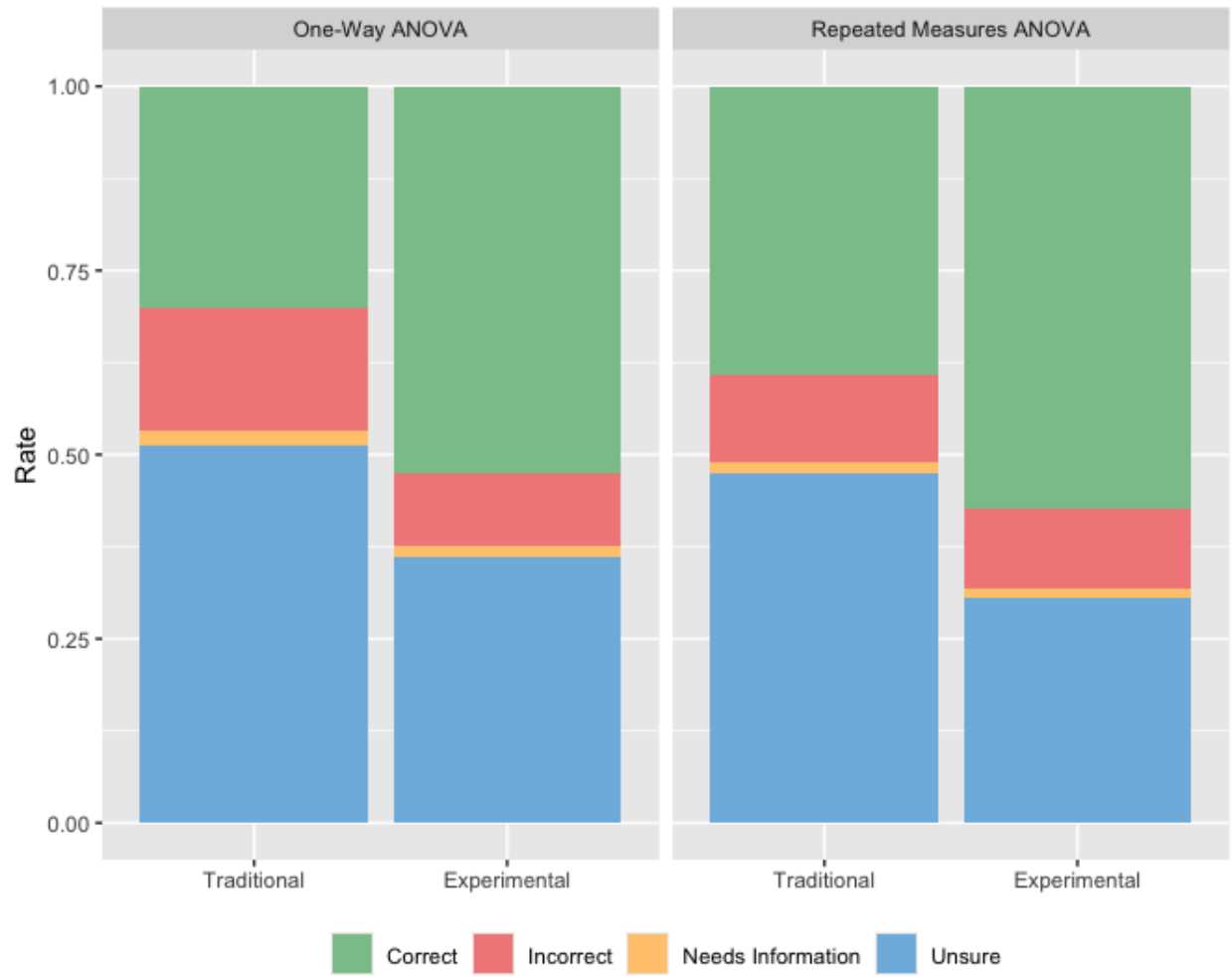


Figure 3.15: Correctness (green) and uncertainty (blue) rates for coding tasks. Overall, the Experimental group showed slightly higher correctness and slightly lower uncertainty across both tasks compared to the Traditional group. This pattern suggests that the Experimental coding syntax supports engagement and improves first-pass accuracy when participants attempt the tasks.

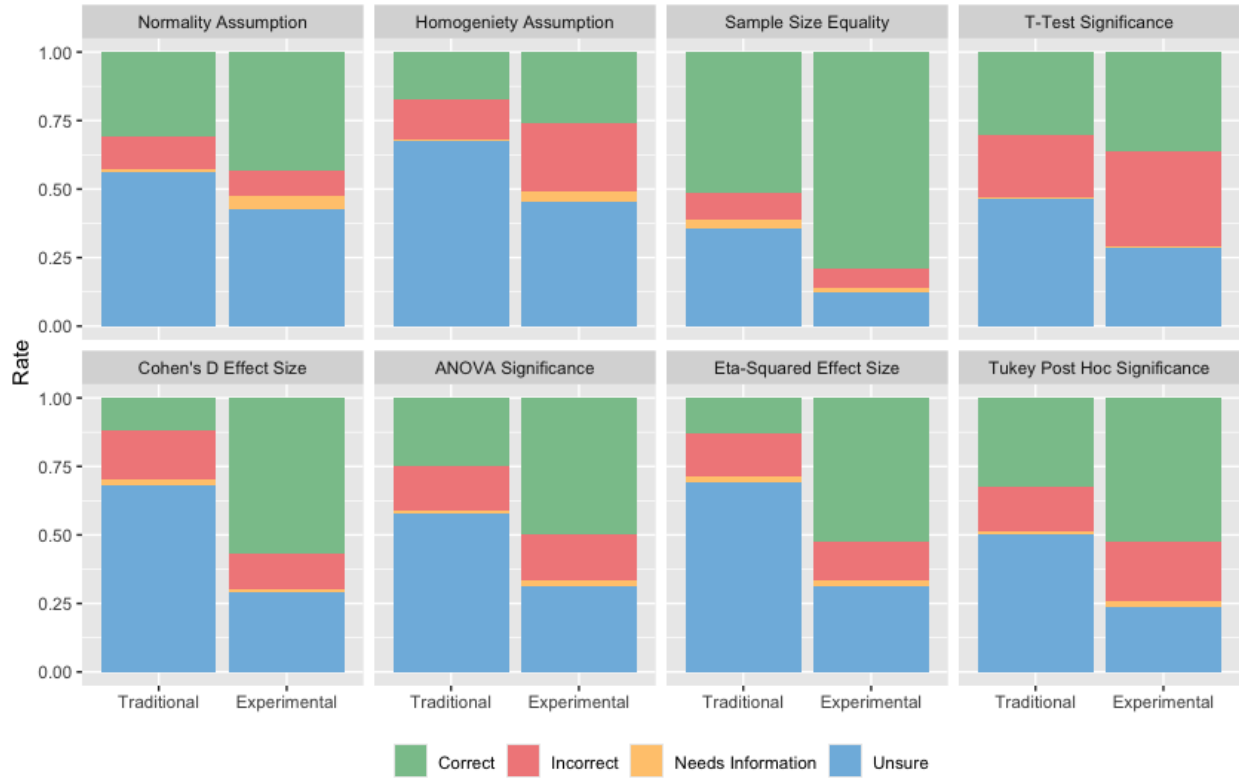


Figure 3.16: Correctness (green) and uncertainty (blue) rates for output interpretation questions. Overall, the Experimental group showed higher correctness across most concepts, while the Traditional group exhibited higher uncertainty. One exception was interpreting t-test significance from a p-value, where correctness was similar across groups. These results suggest that the Experimental output design improves engagement and supports more accurate interpretation of statistical results.

3.6 Qualitative Feedback

Participants were asked 16 open-ended feedback prompts throughout the study, including two for each of the six tasks, as well as four end-of-study prompts about accessibility, anxiety, difficulty, and general comments. For the task-specific prompts, approximately 15–20% of participants responded to each, whereas the end-of-study prompts elicited higher response rates, ranging from 17% (general comments) to 84% (anxiety and difficulty comments).

While these open-ended responses yielded a large amount of qualitative feedback, the primary focus of this study was quantitative. This screening served three main purposes: (1) to identify recurring sentiments that stood out within each design group, and (2) to capture criticisms and issues related to the study structure or design that could inform future research. Past this, (3) to observe whether what participants say matches observations about anxiety, confidence, or correctness compared to our 1-item scales. Representative quotes were selected and organized to illustrate common patterns, highlight contrasting or unexpected remarks, and provide context that may help explain certain results in the quantitative analysis. Note that participant quotes are presented verbatim with minor corrections to spelling, punctuation, and capitalization for readability. Instances of all-caps used for emphasis are retained as written.

We do want to note, however, that identifying themes for a thematic analysis is not a perfect process and our goals were far from wanting to prove themes or other approaches. That said, we were at least loosely guided by Braun & Clark’s influential work on the topic [15], although we show only the final overarching themes and representative quotations for brevity.

3.6.1 Traditional Representative Quotes

We have attempted to loosely categorize the themes we observed stemming from the Traditional group. In each case, we first name the theme of the quotations we observed. Then, we provide the exact quotations below. As one final bit of context before the themes and quotations, several participants recognized the Traditional software as R (mentioned 18 times), suggesting some familiarity. This may have diminished differences versus the Experimental group.

- Cluelessness and apologies:
 - “I am so sorry that I was a participant” or “I felt bad because I could affect your study”
 - “I had no clue. I am very sorry, what seemed difficult... I just [took] one look and ran away.”
 - “I apologize. I understood the questions, I just didn’t know how to answer them”
- Attempting responses:
 - “Unfamiliarity with any of the concepts provided caused more anxiety when attempting to answer”
 - “I attempted to complete this task, but got staggered attempting to complete it”

- “I was very confident that I did not know the answers to any of the questions”
- Hopelessness, apathy, and disengagement:
 - “It was hard for me to gauge my anxiety because I just shut down”
 - “Anxiety went down... because I realized it’s hopeless for me”
 - “I was able to more confidently answer questions from output task 5, but I gave up on the rest”
 - “I felt anxious that I couldn’t answer the questions. Then I gave up and did not feel anxious”
 - “I just kind of gave up, so I felt less anxious”
- Frustration:
 - “I’m not sure if anxiety is the best way to describe what I felt, maybe frustration or bewilderment”
 - “The names do not mean anything to me so that is a bit frustrating”
 - “All forms of anxiousness ... derived from a mixture of fear of not understanding the material and frustration that I can’t complete it.”
 - “I was a bit frustrated to not remember my stat course, rather than anxious.”
- Overwhelm:
 - “The ones with code made me feel overwhelmed and confused”
 - “I also feel very anxious with anything involving code because it overwhelms me”
 - “Overwhelmed by seeing many unfamiliar terms”
 - “I felt anxious because I felt overwhelmed with a bigger mass of data”
 - “There was a lot of data in lump form which made me anxious to look at”
- Confusion about software design:
 - “I do not know why the ‘responses’ variable is created or why it is used during the aov function”
 - “I do not know what any of these functions mean such as aov”
 - “If by ‘the naming convention of the statistical tests’ you are referring to cor t and chisq then yes, those give me anxiety. They’re awful”
 - “The coding portion of this test was what contributed negatively to my anxiety”
 - “The formatting of the output was hard to interpret and understand”
 - “Some of my anxiety did come from how the outputs seemed to get more and more confusing, and their readability seemed to get worse and worse.”

3.6.2 Experimental Representative Quotes

Similarly, we provide here a good-faith attempt at reasonable themes we observed in the experimental group. Again, the theme is provided, along with representative quotations.

- Apologizing:
 - “I don’t really know how to explain, I’m sorry, but 3 and 4 were the most stressful”
 - “Sorry for all the spelling mistakes”
- Attempting responses:
 - “Gave me a sense of ease as I could either answer them or at least make an attempt”
 - “I had no anxiety attempting this”
 - “I felt uncomfortable when attempting these questions”
- Hopelessness, apathy, or disengagement:
 - “After the first two, I completely gave up and [didn’t] feel anxious anymore”
 - “The questions I did try on, I felt anxious but if it looked too hard I just put unsure to avoid any anxiety”
 - “I just gave up on some questions and didn’t let it bother me”
 - “Some tasks felt too difficult, so I didn’t engage fully with them”
- Frustration/negative self-perception:
 - “The anxiety has to do with my frustration of being unfamiliar with this type of coding”
 - “Coding is just so scary to me, I don’t know why but it makes me feel stupid”
 - “I felt anxious, because I have always been bad at math, and seeing numbers already gave me anxiety”
- Overwhelm (less negative especially in regards to the study rather than statistics or math):
 - “The table helped organize information so I could focus on it without getting overwhelmed on the page”
 - “Stats has always been slightly overwhelming to me”
- Confusion about software design:
 - “CompareCounts() <- COUNTS? OF WHAT???”
 - “I wasn’t sure what ‘correlate groups’ meant”

- “I have no idea what `SelectRepeatedFactor()` really does”
- “Not knowing exactly what the ‘`SelectRepeatedFactor`’ line means made me less confident”
- “The output tables in tasks 5 and 6 seemed to make it easy to answer the questions but that easiness led to feelings of confusion”

Different readers will come to different conclusions on the overall meaning of the quotations, but our takeaway is that the qualitative feedback seems inline with the quantitative results. The Traditional group had a large amount of negative feedback, which felt less in the experimental, although this was difficult to quantify in a way we found satisfactory. Further, areas where the quantitative results indicate the experimental group still needed improvements to its design (e.g., `CompareCounts`) directly showed up in qualitative feedback, indicating a form of match. If anything, it told us how people felt when doing the study and hit home with our team that anxiety and other issues in statistics are not just theoretical. Design and statistical anxiety impacts real people, including researchers and academics.

3.6.3 Study Format Feedback

Participants also raised points about the study’s format itself. A few noted that the white background of tasks made them “nervous” or made it hard to “focus.” The sub-scale wording “How much did each of the following contribute to your anxiety/confidence” was confusing to some, who were unsure whether it referred to positive or negative contributions.

Several participants mentioned being confused by certain questions or instructions, with two criticizing the top-to-bottom numbering rather than left-to-right, two noting irritation with scrolling, and four requesting a zoom function. One participant said the dataset snippets were too short to depict intended variable types, while another criticized the realism of the apple dataset (“Maybe I’m overthinking this a bit. Perhaps. Haha”). One was upset that the text-box for code did not allow tabbing, making code “ugly.” Many participants in both groups wished they could test code in real time. This final comment highlights one of the limitations in a study like this one, which is that practice over time matters and may very well impact both correctness and how people feel about statistics.

Accessibility was also raised: multiple Traditional group participants said the output would be difficult for those with dyslexia (one of which self-reported as dyslexic), while another participant said the study’s screen reader accessibility worked well. Finally, some remarked that the study felt like a “test” or “exam” that created worry about doing well, while others noted they did not feel anxious since they knew there were no formal stakes or grades.

3.7 Discussion

3.7.1 Primary Findings: Software Design Effects

The Experimental software design consistently improved users' experiences and task performance compared with the Traditional design, showing how interface choices can shape whether people can answer questions about statistics. Past this, it can also impact both affective responses and engagement. Correctness gains were largest in programming input (+14.5%) and output interpretation tasks (+9.1%), with smaller and mixed effects in selection tasks. In some test-specific cases, the Experimental group performed worse, highlighting the need for careful consideration of descriptive *and* intuitive naming conventions for each unique statistical method. It also means "say what you do" is generally a theory supported by the data here, but it either is not universal or there are gaps yet to explore.

Participants also reported lower uncertainty on all tasks, selection (−38.2%), input (−15.9%), and output (−19.9%), which may mean that the simpler language and syntax encourages active engagement rather than disengagement or "I am not sure" responses. Confidence improved consistently (+4.1% overall), suggesting modest interface changes can boost self-efficacy and facilitate decision-making. Anxiety reductions were smaller and task-specific (−3.3%), primarily in input and output tasks, demonstrating that design interventions can help alleviate affective barriers. The impact is not very large, which means it may not be detectable in future studies if they have small samples.

Perceived difficulty decreased in the more cognitively demanding input (−11.4%) and output (−11.2%) tasks, but remained largely unchanged for selection tasks, suggesting that interface improvements are most effective where reasoning demands are highest. Overall, these findings suggest that even modest interface improvements can make computational and statistical tools more approachable for learners, potentially improving cognitive and affective outcomes, and supporting skill development in computing education.

Taken together, these findings provide the first empirical evaluation of the "say what you do" design principles. The evidence so far is in support, but leaves many unanswered questions. The changes to even simple words translated from terminology preference into measurable improvements in anxiety, confidence, and correctness under a variety of metrics and tasks.

3.7.2 Secondary Analyses and Broader Patterns

Gender and Field Effects. Gender differences in perceptions were generally small but somewhat consistent with prior research. Women reported slightly higher anxiety and lower confidence on some tasks, but these differences explained only 1 to 4% of the variance and did not extend to performance, which was equivalent across genders. Much of the difference in genders appeared to be related to the data- and information-related concepts, supporting Burnett's work regarding women's lower self-efficacy [8]. This pattern also aligns with prior studies showing that women often report greater statistics anxiety and lower confidence, even when objective outcomes are comparable [114, 112, 85, 134]. However, these gender differ-

ences were often shaped by academic field: women in the formal sciences reported higher confidence relative to women in other fields, while men’s confidence remained stable across domains. Notably, women in formal sciences often had comparable anxiety and confidence to that of men. This supports findings from Kian [68] and Mji [95] that gender effects can be small or reversed in different domains, and Trassi’s [134] review noting prior samples often over-represent women in social sciences, potentially inflating observed gender gaps.

Relationships Among Perceptions. The Correlation was moderate and negative between anxiety and confidence ($r = -.29$), consistent with earlier work linking heightened anxiety to reduced self-efficacy [104, 87, 47], and this association was consistent across each of the task areas. Confidence was also strongly associated with uncertainty ($r = -.49$), consistent with Macher [87], Kaufmann [66], and Hoegler [59], who linked confidence to greater engagement and persistence in statistical problem-solving.

However, anxiety showed weaker and more inconsistent relationships with uncertainty ($r = .19$). This suggests that confidence-building interventions may be more impactful than generalized anxiety reduction. Open-ended feedback reflected this, as several participants said they did not feel anxious because they simply “gave up,” and one Traditional group participant reported using unauthorized resources due to anxiety, echoing Sandoz’s [119] and Eshet’s [40] observations linking statistical anxiety to negative coping strategies such as academic dishonesty or avoidance. These patterns suggest that confidence is a primary affective driver of engagement and willingness to attempt tasks, with important implications for reducing avoidance, uncertainty, and negative coping in statistics learning.

These findings also raise a methodological point worth acknowledging. The single-item measures used here are clearly not universally appropriate, contexts such as classroom instruction or settings where software is not central to the task may call for different approaches. That said, the measures employed here demonstrated acceptable test-retest reliability, convergent validity with existing literature, and consistency with qualitative feedback. Their exact boundaries remain an open question, and future work should explore alternative measurement strategies, particularly for capturing affective responses to statistical software in varied learning environments.

Perceived Difficulty Across Domains. Task difficulty ratings varied systematically across domains, with output (interpretation) rated the most difficult, followed by input (programming), and selection was rated as the easiest. This pattern explained 34% of the variance and is consistent with prior work identifying interpretation as especially anxiety-inducing [101, 86, 120] and with Lytsy [83], who found difficulty persisted even for trained individuals. Similar comprehension issues have been documented in peer-reviewed literature [2, 57]. The results confirm this ongoing interpretation difficulty even for trained individuals, as participants with more statistical experience were more engaged but did not achieve higher accuracy. In this study, the experimental software design mitigated some of the difficulty associated with interpretation tasks, underscoring the potential of design-based interventions to reduce cognitive load and support understanding in high-demand domains, this builds on the previous software-related findings [115, 111, 47, 62].

Disability Status. To our knowledge, this study is the first to examine statistical anxiety and performance in relation to disability status in this context. Participants reporting one or more disabilities experienced significantly higher statistical anxiety, particularly during the test-selection and interpretation tasks. To our knowledge, this is also the first study to systematically examine such differences after ensuring accessibility of the web platform using Web Content Accessibility Guidelines 2.1 AA [141]. In the Traditional interface design group, these participants dropped out at a higher rate than the group average, whereas dropout rates in the Experimental group were lower. This pattern suggests that the redesigned interface may have been more approachable for users with disabilities.

Among those who completed the study, performance and confidence were comparable to participants without disabilities, indicating that, when able to engage, participants with disabilities were equally capable. These effects may have been muted by dropout rates and the wide range of reported disabilities being constrained to only disability status, but they highlight the importance of considering accessibility when designing statistical software or teaching statistical programming.

3.7.3 Uncertainty as a Distinct Construct

A methodological contribution of this study is the explicit treatment of uncertainty as a construct separate from correctness. Participants could select “I am not sure” in response to any question, and these responses were excluded from correctness scoring rather than counted as errors. This may matter: qualitative feedback from the Traditional group consistently revealed that “I am not sure” responses often reflected hesitation, disengagement, or anxiety rather than a genuine absence of knowledge. Several participants described giving up or shutting down rather than attempting answers, responses that, under conventional scoring, would be indistinguishable from wrong answers.

This has implications for how performance is measured in studies of statistical anxiety and software usability. Prior work that scores uncertainty as incorrect may underestimate knowledge while conflating two different phenomena: not *knowing* and not *trying*. The large design effect on uncertainty rates observed here, the Traditional group was 16.2% more uncertain study-wide, and up to 38.2% more uncertain on specific test-selection questions, suggests that interface design influences willingness to engage as much as it influences actual knowledge. Confidence, rather than anxiety, emerged as the stronger predictor of uncertainty, reinforcing the theory that engagement is primarily self-efficacy-driven rather than fear-driven. Future work should treat uncertainty and correctness as separate dependent variables, as collapsing them risks obscuring the affective dimension of statistical task performance.

3.7.4 Interpretation: Where and Why Design Helped

Taken together, the pattern of results suggests that the Experimental design was most effective where cognitive demands were highest and where interface elements most directly scaffolded task performance. Gains in input and output tasks, where participants had to produce or interpret statistical content rather

than simply select from options, were consistently larger than in selection tasks. This is consistent with prior software-related findings suggesting that interface design has greater impact where task demands are highest [115, 111, 47, 62]. In this study, design features such as descriptive naming, structured syntax, and plain-language output appear to have been most beneficial precisely where participants faced the greatest gap between their existing knowledge and the demands of the task. In selection tasks, where the primary challenge was recognizing which test to use, naming conventions influenced engagement and confidence but did not reliably translate to correctness, suggesting that label clarity alone cannot substitute for conceptual knowledge.

The relationship between confidence, uncertainty, and anxiety also points to a mechanism worth noting. Across all task areas, confidence was a stronger and more consistent predictor of both uncertainty and correctness than anxiety was. This suggests the Experimental design worked primarily by raising self-efficacy, making users feel more capable, rather than by directly reducing fear, consistent with prior work linking self-efficacy to engagement and persistence in statistical tasks [87, 66, 59]. The practical implication is that interface interventions may be more effective when designed to scaffold competence than when designed to minimize difficulty. These two goals are related but not identical: a tool can feel easier without making users feel more capable, and the latter appears to matter more for engagement and performance.

Finally, the uneven effects across specific tests, where some naming changes improved both engagement and correctness while others improved engagement at the expense of accuracy, suggest that the “say what you do” principle requires careful, test-specific implementation rather than uniform application. Names like *CompareMeans* succeeded because the label accurately captured both the procedure and its purpose. Names like *CompareCounts* succeeded in drawing engagement but confused users about what was being counted, illustrating that descriptiveness alone is insufficient if the description is ambiguous in context. The theory is thus in support but will need adjustment to flesh out what it does and does not predict and where.

3.7.5 Generalizability

The sample’s diversity across educational backgrounds, academic fields, age groups, and disability status strengthens the external relevance of these findings relative to prior statistical anxiety research, which has predominantly drawn from homogeneous, female-majority, social science populations. At the same time, the generalizability of these findings is shaped by several sample and methodological factors. Most participants were undergraduate and graduate students from two universities, with relatively few academic researchers, working professionals, or individuals from underrepresented demographic groups (e.g., certain disabilities, older adults, or a broader range of academic fields). A small subset of participants were not currently enrolled but were largely former students with prior higher education and statistical experience, and are therefore interpreted within the broader adult learner population rather than as a distinct non-student group. Although the sample included diversity in age, education, and discipline, these subgroup sizes were small enough to require some category consolidation in analyses. This limits the ability to draw strong

conclusions about less-represented populations.

Ecologically, the study’s online, self-paced format allowed participation in varied settings, improving accessibility and reflecting how statistical tools are often used in practice. However, it also introduced uncontrolled variability in distractions, motivation, and available resources. Recruitment timing, primarily at the start of the semester, meant that many participants completed the study later, sometimes after major exams, which may have influenced their performance, confidence, and anxiety.

The statistical tasks simulated real-world decision-making and interpretation but used simplified datasets and tightly scoped procedures (Pearson correlation, chi-square, t-test, and one-way/repeated-measures ANOVAs). This focus supports controlled comparisons but limits applicability to more complex, messy, or domain-specific analyses (e.g., regression models, multivariate methods, non-parametric alternatives). Selection and output questions were multiple-choice, which could invite guessing; however, explicit “I am not sure” and “There is not enough information” options were provided to discourage random selection.

The programming component was implemented with realistic coding requirements but in a non-runnable, text-only interface. This means they could not execute or iteratively debug their code. This contrasts with a standard IDE that catches syntax errors but has no understanding of statistical semantics. For example, a call such as *SetRepeatedFactor("time", "before,during")* missing one repeated measures level, or its equivalent in R, would execute without error and return statistical output with no indication that the repeated measures structure was misspecified [140]. Point being, many aspects of statistics are errors at runtime, so this limitation is important but not always.

Our rubric explicitly separates syntactic from conceptual criteria (see Tables 3.5 and 3.6), allowing evaluation of whether participants understood the statistical structure of each problem independently of whether their code would parse correctly. It is also worth noting that each programming task provided participants with an explicit worked example using a different dataset, requiring only variable substitution to complete, a scaffold that reduced syntactic demands but does not reflect the full complexity of real-world statistical programming. In a more realistic setting, data formatting, variable management, and task framing are rarely provided in advance.

An additional limitation concerns the one-time scope of the study. Single-session performance measures do not capture how users engage with statistical software over time, and this distinction is worth considering. Unlike general programming, which may be practiced regularly, statistical software may only be used when a specific analysis is needed. In these contexts, users might need to relearn conventions each time they return to the tool, a pattern that is arguably more common in computing education research, where statistical analysis is just one of many demands on a researcher’s time. Studies like this one provide hints and early evidence, but far more work is needed.

The fact that participants still showed substantial variability in correctness under these conditions suggests the tasks were sensitive to genuine conceptual differences rather than surface-level coding familiarity. That said, performance measures here are strictly constrained to first-attempt reasoning and do not capture

iterative refinement processes common in practice. Future work should explore environments that go beyond plain text but also beyond standard IDEs, tools capable of guiding users on both syntactic and semantic errors, to more fully capture the range of performance improvements that interface design may support.

Another limitation concerns the measurement of key constructs. Anxiety and confidence were assessed using single-item Likert scales (7-point), while end-of-study difficulty was measured with a 5-point scale inspired by established difficulty measures [136]. The rationale for and psychometric evaluation of these single-item measures is discussed in detail in Section 3.5.3; briefly, established multi-item instruments were not feasible for repeated task-level measurement, and pilot testing of a two-item scale revealed reliability issues that motivated the final approach. While these measures demonstrated acceptable test-retest reliability, convergent validity, and largely matched the qualitative responses, what statistical anxiety even “is” is easily of debate.

For example, established multi-item instruments are designed to measure trait-level statistical anxiety and self-efficacy as stable dispositions, whereas the present study required task-specific measurements of affective responses during moment-to-moment software interaction. A construct that the established instruments do not mention and were not designed to capture. That said, the single-item scales lack the psychometric depth of multi-item scales [139, 29, 56, 44] and may not fully capture the complexity of statistical anxiety and self-efficacy as trait-level constructs.

Additionally, we note that the Kincaid reading level calculation [71, 70] has important limitations when applied to statistical software output. The metric is based solely on sentence length and number of syllables per word, which makes it appropriate for prose but less suitable for text containing code syntax, abbreviations, variable names, or numerical results. Accordingly, in this study we used the Kincaid score only as a rough guide to simplify the wording and conceptual language of textual output. Numerical values, symbols, and code fragments were not factored into readability judgments.

In sum, while this study has limitations related to subgroup power, measurement depth, procedural scope, family-wise error in secondary analyses, and the constrained coding environment, the consistency of findings across conditions, outcome measures, and demographic groups provides convergent evidence for the benefits of user-centered statistical software design. These limitations also point toward a clear agenda for future work, several directions for which are outlined in the following section.

3.7.6 Implications

The results of this study are particularly relevant for three overlapping audiences: instructors selecting statistical software for courses, curriculum designers integrating statistical tools into computing and statistics programs, and developers designing statistics tools. We organize the practical implications of our findings below by audience.

Implications for Instructors Instructors making decisions about which tools to adopt may benefit from considering not only functionality, but also how software design influences students’ anxiety, confidence, and ability to interpret results. Statistical software should be viewed as part of the instructional environment rather than a neutral tool: design choices such as clearer test naming, consistent and structured input syntax, and more interpretable output may reduce cognitive load and support understanding, particularly in areas like test selection and output interpretation where confusion is common. These implications extend across instructional levels, from undergraduate introductory courses where statistics anxiety is most acute, to graduate programs where statistical software is used routinely in research contexts but formal training may be limited. Such considerations are especially relevant for learners from non-STEM backgrounds or those with lower baseline programming skills.

Implications for Curriculum Designers As data literacy becomes essential across disciplines, curriculum designers integrating statistical computing into CS and statistics programs can use human-centered software design to help shift statistics from a gate-keeping subject to an accessible, empowering skill set. Institutions might consider integrating tools with reduced cognitive barriers into curricula to lower entry barriers, support confidence, and reduce anxiety, especially for underrepresented groups or those with prior negative experiences in statistics. Embedding such tools with real-world relevance but reduced cognitive barriers *could* improve retention, engagement, or attitudes toward statistical work.

Implications for Software Developers For developers designing statistics tools, our results suggest that modest interface changes, such as clearer code, reorganized output, and accessible test names, can improve confidence, reduce uncertainty, and in some cases boost performance and lower anxiety, especially in high-demand areas like programming and interpretation. These design considerations have implications for both educational and professional contexts. Making output clearer may also assist the general public in understanding results in some cases.

Future research should expand participant diversity and task scope to include regression, data preparation, and visualization, testing whether design benefits persist in more complex workflows. More broadly, these findings highlight that statistical software design is not merely a technical concern, but an educational one, shaping how learners engage with and develop confidence in statistical reasoning.

Chapter 4

Study 2: Learning to Swim in a Sea of Statistics Anxiety: A Randomized Controlled Trial

4.1 Introduction

Building on the *say what you do* naming theory, which demonstrated that descriptive statistical function names are strongly preferred over conventional names [72] and improve user engagement (see Study 1 in Chapter 3, Figure 3.14). The first study investigated whether broader surface-level simplifications to statistical workflows could further reduce barriers to applied analysis. Modifying test names, reducing syntactic complexity, and providing plain-language output improved confidence and engagement and slightly reduced statistical anxiety. Yet analytical accuracy showed mixed results, and anxiety persisted. These findings suggest that surface simplification alone is insufficient: many difficulties in applied statistics stem from the procedural and structural complexity imposed by the tools themselves, not from conceptual misunderstanding. Importantly, participants demonstrated a strong ability and willingness to identify variable types, roles, and structural relationships within their data (see Figure 3.11), indicating that users may understand their datasets better than the tools allow them to express. This led to the present study, introducing *Declarative Data Science*, a semantic, structure-first approach in which users explicitly declare variable types, roles, and relationships once, after which procedural decisions follow deterministically. This relocates cognitive effort away from remembering test names, argument orders, or encoding rules, and toward articulating the structure of the research design and data itself.

This work is also guided by earlier structural abstractions in statistical computing. A symbolic formula notation was developed in 1973 by Wilkinson and Rogers [148] and first implemented in 1991 in the language

S, R’s predecessor, advancing the field by allowing users to specify variable relationships rather than invoking algorithms directly. However, it leaves critical semantics implicit: variable type distinctions, repeated or hierarchical structure, and encoding behavior are inferred or expressed through syntax convention, increasing the likelihood of hidden assumptions and silently incorrect analyses. Wrapper packages such as R’s *ez* [79] and Python’s *pingouin* [137] reduce surface inconsistency but do not eliminate the cognitive burden of test selection, and as analytical complexity increases, their design constraints often require users to fall back on formula-based frameworks, re-introducing the syntactic complexity they were intended to reduce and imposing multiple learning curves within a single analytical workflow. For machine learning, Python’s *scikit-learn* [107] standardizes model interfaces and offers pipeline tools to chain preprocessing steps, but the user must still explicitly specify which transformations apply to which variables, determine appropriate encodings and scaling per model type, and construct the full preprocessing specification manually. Across all three approaches, semantics that cannot be inferred from data alone (variable roles, ordered distinctions, repeated measures structure, covariate status, and ambiguous variable types) are left to the user to remember or express correctly and interpret manually. Any omissions or mistakes can produce not just errors but also silently incorrect results.

4.1.1 What is Declarative Data Science?

Statistical computing tools have historically followed an imperative workflow: the user supplies the full procedure, selecting an algorithm, assigning variables to argument slots, and applying preprocessing and typing manually. Wilkinson notation represents a partial step toward declarative specification; the user describes variable relationships symbolically using a tilde to separate outcomes from predictors, $+$ to add terms, and $*$ to specify interactions. However, this symbolic structure carries far less semantic information than it may appear. The formula $y \sim x1 * x2$ does not specify whether $x1$ is continuous or categorical, whether it represents a between-subjects factor or a within-subjects repeated measure, or whether it is a primary predictor or a covariate being controlled for. *Declarative Data Science* addresses this gap by requiring the user to fully specify the semantic structure of their analytical problem before any procedural decision is made. Every variable is assigned a *type*: *Numerical* (treated as continuous), *Ordered* (treated as ordinal), or *Categorical* (treated as nominal); and a *role*: *Input* (a between-subjects predictor), *Repeated* (a within-subjects factor), *Control* (a covariate of secondary interest), *Output* (the outcome variable), or *Identifier* (links repeated observations). Relationships between kinds of input variables, analogous to interaction terms in Wilkinson notation, are declared explicitly rather than inferred from formula syntax. Together, these declarations fully specify the analytical design and the appropriate test or model, encoding strategy, and preprocessing steps can follow deterministically, with standard defaults applied unless the user specifies otherwise. Two decisions remain intentionally with the user: 1) the *analytical goal*, what kind of question is being asked of the data, such as inference from existing data (*GetTest*) or training a predictive model for future observations (*GetModel*), and 2) the *semantic structure* of the problem itself. Both are intrinsic to

understanding one’s research question regardless of software. Everything else follows automatically. This separates analytical meaning from computational implementation, reducing cognitive load while minimizing silent errors, and creates a natural foundation for downstream automation including plain-language output generation and meaningful default visualizations.

This chapter presents the design, methodology, and results of a study evaluating whether *Declarative Data Science* improves statistical reasoning accuracy, task engagement, and affective responses relative to a traditional workflow in which users manually select tests, populate parameters, and handle preprocessing.

4.2 Hypotheses, Aims, and Objectives

This study evaluates the impact of a structure-based, automated data science system on task performance, engagement, and user experience. Building on the first study, which showed that interface simplifications (e.g., clearer test names, simplified syntax, English-language output) can reduce statistical anxiety and improve confidence, this study investigates whether explicitly defining variable roles and types while automating test/model selection can further improve user engagement, understanding, and affective responses across repeated and more complex tasks.

Participants completed a series of 12 statistical and machine learning tasks using either the *experimental* or *control* interface. The experimental interface required users to explicitly define variable types and roles, after which the appropriate test or model was selected automatically. The control interface required users to manually select tests and models, assign variables to parameters, and perform any necessary preprocessing. Primary outcomes include task correctness (first attempt and first intent), task engagement (give-up rate and effort prior to disengagement), and self-reported anxiety and confidence ratings collected after each task.

H₁ Engagement. There will be no difference in task persistence between users of the Experimental and Control interfaces, as reflected in give-up rates, effort prior to disengagement, and frequency of repetitive identical submissions indicative of frustration.

H₂ Correctness. There will be no difference in task correctness between users of the Experimental and Control interfaces, even when considering only those who engaged with the tasks.

H₃ Confidence. There will be no difference in self-reported statistical self-efficacy (confidence) between users of the Experimental and Control interfaces across tasks.

H₄ Anxiety. There will be no difference in self-reported task-related statistical anxiety between users of the Experimental and Control interfaces.

4.3 Method

4.3.1 Inclusion and Exclusion

Participation was restricted to individuals aged 18 years or older. All participants provided informed consent before beginning the study. Two participants indicated their gender as *Other* ($n = 1$) or *Prefer not to say* ($n = 1$). Due to insufficient representation, these participants were excluded from any analyses in which gender was included as a covariate.

4.3.2 Participant Characteristics

Participant characteristics are reported for all individuals who reasonably engaged with the study, comprising completers ($n = 112$) and non-completers ($n = 8$).

Participants ranged in age from 18 to 55 years ($M = 23.47$, $SD = 6.37$) and were predominantly undergraduate students (experimental: $n = 54$; control: $n = 49$), with a smaller graduate and post-graduate representation (experimental: $n = 7$; control: $n = 10$). Highest completed education ranged from high school or GED ($n = 34$) to Ph.D. ($n = 1$), with the majority reporting some college or university experience ($n = 48$) or a bachelor’s degree ($n = 17$). All participants were enrolled in or had completed STEM degree programs, predominantly Computer Science ($n = 73$, 60.8%), Computer Engineering ($n = 19$, 15.8%), and Cybersecurity ($n = 12$, 10.0%). The homogeneity of the sample is noted as a generalizability limitation.

Prior statistics coursework was limited; 52 participants (43.3%) reported no course experience and 51 (42.5%) reported one course. Coding experience was moderate with an average of about 4 courses and machine learning experience was minimal (less than one course on average). Proficiency ratings (1–5 scale: *None* to *Expert*) reflected this pattern: statistics fell between Basic and Intermediate, and coding between Intermediate and Advanced. Machine learning, Python and R were all Basic, while Quorum was close to None.

English was the primary language of 104 participants (86.7%); among the remaining 16, Spanish was most common ($n = 6$), followed by Amharic ($n = 3$), Chinese ($n = 2$), and French, Thai/Laotian, and Vietnamese ($n = 1$ each). Twenty-two participants (18.3%) disclosed one or more disabilities, including attention deficit disorders ($n = 12$), mental health conditions ($n = 8$), visual impairment ($n = 3$), autism ($n = 4$), mobility impairment ($n = 2$), and other conditions ($n = 2$).

4.3.3 Sampling Procedures

Responses were collected online over a two-month period from March 2026 to May 2026. Participants self-selected into the study through an online platform hosted on a secure university server. To control for gender and maintain balanced groups, participants were randomly assigned to one of the two software versions of the study based on their response to the gender question.

Table 4.1: Participant demographic characteristics (N = 120)

Characteristic	Experimental (n = 61)	Control (n = 59)
<i>Age (Mean (SD))</i>		
Years	23.03 (5.80)	23.92 (6.92)
<i>Gender</i>		
Female	18	17
Male	43	40
Other	0	1
Prefer not to say	0	1
<i>Current Education Level</i>		
Undergraduate	54	49
Graduate / post-graduate	7	10
<i>Primary Language</i>		
English	52	52
Other	9	7
<i>Prior Experience (number of courses, Mean (SD))</i>		
Statistics	0.64 (0.78)	0.88 (0.89)
Coding	3.97 (3.62)	4.36 (3.39)
Machine Learning	0.46 (0.98)	0.53 (0.88)
<i>Self-Rated Proficiency (1–5 scale, Mean (SD))</i>		
Statistics	2.31 (0.85)	2.42 (0.95)
Coding	2.95 (0.76)	3.25 (1.03)
Machine Learning	2.00 (0.88)	2.03 (1.17)
Python	2.23 (0.92)	2.34 (1.20)
R	1.57 (0.74)	1.54 (0.75)
Quorum	1.31 (0.53)	1.24 (0.47)
<i>Disability Status</i>		
Disclosed one or more disabilities	10	12

Recruitment occurred through advertisements sent to over 70 university professors in STEM departments. In-person recruitment took place in several courses per professors' requests. When applicable, participants were offered 2% extra credit for participation if approved by their professors, with alternative avenues provided for those who chose not to participate. Once the study concluded in May 2026, professors who had approved extra credit received a list of students who participated and provided their student email. Student emails were discarded after extra credit was issued. The study was approved by the Institutional Review Board (IRB) at the University of Nevada, Las Vegas (UNLV), adhering to ethical standards of informed consent, confidentiality, and voluntary participation.

A total of 141 individuals completed the demographic questionnaire. Several categories of exclusion were applied to removed non-viable data, as described below. In total, 112 participants completed the entire study and were retained for full analyses, with 8 additional participants providing enough data for partial analysis.

Non-starters. *Excluded.* Participants who completed the demographic questionnaire but did not begin Task 1 were excluded, as their data contains no task-level information relevant to the research questions ($n = 9$; 2 did not engage with the practice task, 7 completed the practice task but did not proceed to Task 1).

Non-engagers. *Excluded.* Participants who completed the study procedurally but demonstrated no meaningful engagement were excluded ($n = 8$; 3 experimental and 5 control). At an average of fewer than 25 seconds per task, the total study duration of under five minutes is physically insufficient to read task descriptions, interact with the interface, and provide thoughtful responses. Exclusion was supported by converging evidence: five of eight participants skipped portions or the entirety of the practice task, none provided qualitative feedback across any task, all made essentially zero attempts before immediately giving up on every task, and affective ratings displayed flat or random patterns inconsistent with genuine task engagement.

Technical dropouts. *Excluded.* Despite prominent instructions displayed on every page explicitly warning participants not to use browser navigation controls or reload the page, four participants did so (2 experimental and 2 control), corrupting their session data. In two cases, participants contacted the researchers directly to apologize and requested to resume the study. However, as their data could no longer be reliably tracked or attributed to a single session, these participants were excluded on data integrity grounds regardless of intent, still receiving extra credit where applicable.

Non-completers. *Included.* Eight participants began the study but did not complete all 12 tasks (3 experimental and 5 control). The reasons for non-completion cannot be determined; however, their partial data was retained for descriptive analyses up to the point in which it is available. That said, any analyses requiring complete within-subjects data across all 12 tasks necessarily excludes these participants.

Completers. *Included.* The remaining participants, 112 in total, completed the entire study and were used for all analyses.

4.3.4 Sample Size, Power, and Precision

Sample size was informed by a pilot study ($N = 30$) conducted prior to the main study to iteratively refine the study instrument. Quantitative responses and qualitative feedback from the pilot led to refinements in task ordering to ensure a steady increase in difficulty, feedback priority ordering, interface element placement (buttons, blocks, and scroll bars) to avoid mis-clicks, practice task design to optimize interface familiarity, visual feedback and color aspects to align with expected intent, task descriptions and datasets to reduce ambiguity, and wording of rating scales and demographic items. The interface was additionally refined through iterative internal testing with colleagues prior to and alongside the pilot. Pilot participants were excluded from the main study sample. In the final sample, 61 experimental and 59 control participants were retained, with 58 experimental and 54 control participants completing all 12 tasks.

4.3.5 Measures and Covariates

4.3.5.1 Primary Measures

Correctness. Two pre-feedback correctness metrics are used. *First-attempt correctness* captures the accuracy of the workspace at the moment of first submission. *First-intent correctness* captures the accuracy of each individual reasoning decision the first time it is submitted, accumulated across submissions until all elements have been evaluated unless preempted by a give-up, distinguishing statistical reasoning quality from solution-level execution or hint-reliant performance.

Give-Up Rate. Task disengagement was estimated using two complementary measures. The *raw give-up rate* is the proportion of participants per group who gave up on each task, reported per task to characterize disengagement across increasing complexity. However, raw give-up rate treats all abandonment equivalently regardless of how much effort preceded it. Our proposed *effort-weighted engagement* metric attempts to address this by capturing the full spectrum of engagement behavior on a continuous scale from -1.0 to 1.0 , computed as shown in Equation 4.1, where n is the number of non-consecutive unique submissions.

$$\text{attempt-weighted engagement} = \begin{cases} \frac{-1}{1+n} & \text{unsuccessful outcome} \\ \frac{1}{n} & \text{successful outcome} \end{cases} \quad (4.1)$$

A participant who gives up immediately without any attempt receives a value of -1.0 ; one who succeeds on the first attempt receives 1.0 . As the number of attempts increases, engagement for both give-ups and successes converge towards 0 . This design choice reflects the theoretical distinction between the decision to initiate engagement and the effort exerted once engaged [6]. By construction, the metric weights the initial decision to attempt (or abandon) a task more heavily than subsequent persistence. Similarly, a value of 1.0 is only possible when a participant succeeds on their first submission without having received any corrective feedback, weighted more heavily than feedback-guided success. Consecutive identical submissions

are collapsed into a single attempt to prevent frustration clicks from inflating effort scores. The weighted measure serves as the primary disengagement metric.

Frustration Rate Frustration was operationalized as occurring when a participant made three or more consecutive identical submissions during the task, or two or more consecutive identical submissions immediately followed by giving up without modification. Due to the qualitatively distinct outcomes that can follow a frustration bout (e.g., recovery and success, recovery and eventual give-up, or immediate give-up), frustration is reported descriptively as the proportion of participants per task triggering at least one frustration bout, alongside the proportion of that subset who gave up immediately following the bout.

Anxiety. Measured after each task using a single-item 7-point Likert scale, ranging from *No anxiety at all* (1) to *An extreme amount of anxiety* (7). A single-item format was selected over validated multi-item instruments such as the 51-item STARS [29] to capture real-time task-specific affect without inducing response fatigue across 12 tasks. To anchor responses consistently, participants were provided the following definition prior to rating: “*Anxiety is a feeling of fear or apprehension when presented with statistics content and problems or when gathering, processing, and interpreting data.*”

Confidence. A self-efficacy metric, measured after each task using a single-item 7-point Likert scale, ranging from *No confidence at all* (1) to *An extreme amount of confidence* (7). As with anxiety, a single-item format was preferred over instruments such as the 14-item CSSE and SELS scales [44], which capture Bandura’s foundational work on self-efficacy theory [7] in a statistics-specific context. This was done to maintain ecological validity within a repeated-task paradigm. Participants were provided the following definition prior to rating: “*Confidence is the sense of one’s own ability to understand and complete actions required to accomplish specific statistics-related tasks.*” As a convergent validity check on these two single-item measures, the relationship between anxiety and confidence ratings was examined using a repeated measures correlation across all tasks and participants; a negative relationship was expected, consistent with prior literature [108, 134, 87, 66].

4.3.5.2 Secondary Measures

Qualitative feedback was collected via an optional open-ended text field after each task, prompting participants to reflect on strategies used, sources of confusion or help, and reasons for their anxiety and confidence ratings. An additional open-ended response was collected at the end of the study reflecting on their overall experience. Selected responses are used to contextualize and illustrate quantitative findings.

Exploratory Analyses Gender and disability status were examined as exploratory grouping factors given documented gaps in statistics anxiety and confidence across gender and the underrepresentation of disabled

users in statistical anxiety or tool research. Due to small subgroup sizes, these analyses are descriptive and interpreted with caution.

Measures Collected But Not Analyzed Age, education level, enrollment year, primary language, academic major, prior coursework, and self-rated proficiency across six domains were also collected. No significant between-group differences were observed at baseline (all $p > .05$); self-rated coding proficiency showed a marginal trend toward higher ratings in the control group ($t = -1.84$, $p = .068$).

4.3.6 Data Collection

Data was collected via an online platform hosted on a secure university server. Session tracking and email validation prevented duplicate entries, ensuring each participant contributed one response. Participants were randomly assigned to the experimental or control group based on their response to the gender demographic item to maintain balanced gender distributions across groups. Following study completion, extra credit was distributed to eligible participants through their instructors in accordance with IRB approval, after which all identifying information including email addresses was permanently removed from the dataset.

4.3.7 Quality of Measurements

Two features of the data collection process support the quality of the measures reported in this chapter. First, all data were collected and scored automatically by the study system itself rather than by human raters or coders. Correctness, frustration, and give-up behavior were recorded directly from participant interactions with the interface and scored deterministically against predefined criteria (Section 4.3.10), eliminating the need for inter-rater reliability assessment, as no human judgment was involved in data collection or scoring. Second, the study collected multiple independent measures, correctness, engagement, frustration, and self-reported affect, from the same participants across the same 12 tasks, providing an opportunity for convergent evidence across measurement approaches.

4.3.8 Instrumentation

Two similar interfaces were implemented as controlled web-based environments accessible on desktop computers in which all syntactic variation was constrained through dropdown menus and pre-built blocks. Neither interface required participants to type code. The manipulation was therefore limited to the conceptual paradigm, variable-first classification in the experimental group versus test-first selection in the control group, rather than differences in coding modality, interface complexity, or syntactic demands.

Each of 12 tasks presented participants with four major components, shown in Figure 4.1. A) A task description panel and B) a scrollable dataset table of fake specifically-crafted data for each scenario identical across both interfaces. C) A scrollable block palette, the primary point of difference between interfaces, was divided into a variable section and a test/model section, with available blocks differing by group as described

in the sections below. D) A workspace area was shared in structure across both interfaces, divided into a variable zone and a test/model zone, limiting the placement of related blocks. The workspace displayed surrounding static code (package imports, dataframe setup) to provide syntactic validity to the blocks placed within it.

A feedback bar above the action buttons displayed “Click ‘Check Answer’ to check your progress” at the start of each task, updating to reflect the system’s evaluation upon each submission. A “Give Up” button and a “Check Answer” button were always visible; the Give Up button was positioned to the left and the Check Answer button to the right. Participants could show or hide the dataset and workspace panels using toggle buttons. Both interfaces included the instruction: “*You are not being tested; do not seek any outside assistance, simply answer to the best of your ability.*” With every page, participants were warned in bold, red font not to use browser navigation or reload buttons during the study.

Tasks. The participants were given 12 tasks to complete using the interface described above. The tasks were ordered to introduce statistical concepts and increase complexity incrementally (Table 4.2). Early tasks (1–3) introduced basic variable types and group comparisons, middle tasks (4–8) layered in interactions, repeated measures, and multiple outcomes, and later tasks (9–10) combined all five variable roles via covariates. The final two tasks (11–12) shifted from hypothesis testing to predictive modeling. A full explanation of each task, dataset, description, rationale, and the correct answer for both interfaces can be found in Appendix B.1.

Table 4.2: Statistical methods and key concepts across the 12 tasks.

Task	Method	Key Concept
1	Pearson correlation	Relationship between two numeric variables
2	One-way ANOVA	Between-subjects factor, a numeric outcome
3	Chi-squared test	Two categorical variables, no numeric outcome
4	Two-way ANOVA	Second input factor and interaction term
5	One-way repeated measures ANOVA	Within-subjects factor and subject identifier
6	Kruskal-Wallis test	Ordinal outcome variable
7	Mixed-design ANOVA	Combined between- and within-subjects factors
8	MANOVA	Multiple outcome variables
9	Two-way ANCOVA	Covariate (control variable)
10	Mixed ANCOVA	All five variable roles combined
11	Linear regression	Predictive modeling; numeric outcome
12	Logistic regression	Predictive modeling; categorical outcome

Practice Task. Prior to the 12 tasks, participants completed an interactive practice task using their assigned interface. The practice task introduced the core interface mechanics: how to drag variable and

Please do NOT use your browser's back, forward, or reload buttons while completing this task.

You are not being tested, do not seek any outside assistance, simply answer to the best of your ability.

You can attempt as many times as you want. If you get stuck, you can click "Give Up" to move to the next task.

Task 2 of 12

Task Description

A teacher wants to evaluate differences in exam performance across study methods. She separates her students into three groups: one uses flash cards, one completes a practice exam, and one receives no study material. Use the dataset she collected to set up the appropriate analysis to compare exam scores across the groups.

Hide Dataset

student_id	study_method	exam_score
1	Flash cards	81
2	Practice exam	73
3	None	66
4	Flash cards	88
5	Practice exam	78
6	None	71

Hide Work Area

Block Palette

Variable Blocks

Drag or click + to add blocks to workspace.

df[" "] = df[" "]

" ".astype(" ")

Test/Model Blocks

Drag or click + to add blocks to workspace.

pg.corr(x=df[" "], y=df[" "])

↓ Scroll for more ↓

Workspace

```

import pingouin as pg
import statsmodels.api as sm
import sklearn as skl
import pandas as pd

df = pd.read_csv('data.csv')

# Variable setup (optional)

Add one or more variable blocks here

# Test / Model

```

Click "Check Answer" to check your progress.

Give Up Check Answer

Figure 4.1: The full task interface, with task description (A), dataset (B), block palette (C), and workspace (D) labeled. The text content in C and D differ between the Control (shown here) and Experimental interfaces.

test blocks from the palette or use the $[+]$ button to add them to the workspace, how to fill in drop-down options, how to remove blocks using the $[\times]$ button, how to access the information tooltip by hovering over or clicking the $[i]$ icon on each block, how to toggle the dataset and workspace panels, and how to scroll through the full block palette to view all available options. The practice task was not scored and did not contribute to any outcome measures.

Experimental Interface. The experimental interface presented participants with a block palette divided into two sections: variable blocks and test/model blocks. The variable section contained exactly three block types across all 12 tasks: a *CreateSubjectIdentifierVariable* block consisting of one drop-down column selector, an *AddVariable* block, which combined a column selector with a type (*Numerical*, *Categorical*, *Ordered*) and role (*Input*, *Repeated*, *Control*, *Output*) constructor encoded in the function name, and an *AddInputInteraction* block, which combined two variables as seen in *AddVariable* but without the *Output* role option. The column selector in each variable block was constrained to only the variables present in the current task’s dataset, preventing participants from referencing variables outside the analysis. The test/model section contained exactly two block types across all 12 tasks: a *GetTest()* block and a *GetModel()* block, neither of which required parameter configuration, the system resolved the appropriate analysis from the variable design submitted.

Upon submitting an answer, each block in the workspace turned green if correct or red if incorrect, providing granular per-block feedback. A text feedback message below the workspace indicated the nature of the error using a priority-ordered system (e.g., “One or more required variables are missing,” “One or more variable types are incorrect,” “One or more variable roles are incorrect”). Figure 4.2 shows the experimental interface during an active task, illustrating the block palette with constrained column dropdowns, the workspace with per-block color feedback, and the variable type and role constructor options.

Control Interface. The control interface presented participants with a block palette containing one variable block type and 15 test and model blocks, both available across all 12 tasks. The single variable block type was an optional *astype()* conversion block (Python’s method for casting a variable to a different data type, e.g., converting a numeric column to categorical), allowing participants to declare explicit type conversions before analysis; in most tasks the test block’s parameter slots enforced type assignments automatically upon submission, making *astype()* declarations unnecessary. As such, red/green color feedback for an *astype()* block could be misleading so it is intentionally left gray regardless of semantic correctness. For tasks 11 and 12, four additional preprocessing blocks became available in the variable section (*StandardScaler*, *OneHotEncoder*, *OrdinalEncoder*, *LabelEncoder*), bringing the total available variable block types to five.

The entire test block turned green or red as a unit upon submission. The same priority-ordered text feedback system was used. Preprocessing blocks in tasks 11 and 12 each received individual green or red feedback upon submission. Figure 4.3 shows the control interface workspace during an active task after a correct submission, illustrating the information tooltip panel, the test block with whole-block color feedback,

Please do NOT use your browser's back, forward, or reload buttons while completing this task.

You are not being tested, do not seek any outside assistance, simply answer to the best of your ability.
You can attempt as many times as you want. If you get stuck, you can click "Give Up" to move to the next task.

Task 1 of 12

Task Description

A researcher is interested in the relationship between sleep and concentration in college students. He measures how many hours each student slept and their reported concentration levels, and wants to determine whether these two measures are related. Use the dataset he collected to set up the appropriate analysis to determine if the two measures are related.

Hide Dataset

student_id	hours_of_sleep	level_of_concentration
1	4.3	62
2	6.2	70
3	5.1	65
4	7.4	78
5	8.1	82
6	6.7	74

Hide Work Area

Block Palette

Variable Blocks

Drag or click + to add blocks to workspace.

design.AddVariable(frame:CreateSubjectIdentifierVariable(" "))

design.AddVariable(frame:Create " " " ")

design.AddInputInteraction(frame:Create " " " ")

Workspace

```
DataFrame frame
frame:Load("data.csv")

ModelDesign design

// Variable setup (required)

Add one or more variable blocks here

design.AddVariable(frame:Create Numerical " " " ")
Input Variable(" hours_of_sleep ")

design.AddVariable(frame:Create Numerical " " " ")
Variable(" " " ")

UntrainedModel model = frame.GetModel(design)
TrainedModel trained = model.Train()
```

One or more required variables are missing.

Give Up **Check Answer**

Figure 4.2: The experimental interface during Task 1, showing feedback after an incorrect submission. The submitted variable block is highlighted green (correct) and the submitted test block red (incorrect). An additional unsubmitted variable block (gray) is visible in the workspace, and the second dropdown is open showing available role options.

Please do NOT use your browser's back, forward, or reload buttons while completing this task.

You are not being tested, do not seek any outside assistance, simply answer to the best of your ability.

You can attempt as many times as you want. If you get stuck, you can click "Give Up" to move to the next task.

Task 1 of 12

Task Description

A researcher is interested in the relationship between sleep and concentration in college students. He measures how many hours each student slept and their reported concentration levels, and wants to determine whether these two measures are related. Use the dataset he collected to set up the appropriate analysis to determine if the two measures are related.

Hide Dataset

student_id	hours_of_sleep	level_of_concentration
1	4.3	62
2	6.2	70
3	5.1	65
4	7.4	78
5	8.1	82
6	6.7	74

Hide Work Area

Test/Model Blocks

Drag or click + to add blocks to workspace.

pg.corr(x=df[""], y=df[""])

pg.rm_corr(data=df, x=df[""], y=df[""], subject=df[""])

Chi-squared independence tests between two categorical variables. x and y: the variables names.

pg.chi2_independence(data=df, x=df[""], y=df[""])

pg.cochran(data=df, dv=df[""], within=df[""], subject="")

Scroll for more

Workspace

```
import pingouin as pg
import statsmodels.api as sm
import sklearn as skl
import pandas as pd
```

```
df = pd.read_csv('data.csv')
```

Variable setup (optional)

Add one or more variable blocks here

df[""] = df[""]
df[""].astype("numeric")

Test / Model

pg.chi2_independence(data=df, x=df[""], y=df[""])

The test or model does not match the task.

Give Up

Check Answer

Figure 4.3: The control interface during Task 1, showing feedback after an incorrect submission. The variable block remains gray (*astype* variable blocks do not change color in the control interface) and the test block is highlighted red indicating an incorrect submission. The information tooltip for the chi-squared test block is open, displaying a brief description of the test.

Table 4.3: Experimental interface block palette information tooltip provided.

Block	Information Tooltip
<i>Variable Blocks (Shown in all tasks)</i>	
<i>SubjectIdentifier</i>	Add a variable that identifies which rows belong to the same individual to your design.
<i>AddVariable()</i>	Add a variable to your design by specifying its data type and role. INPUT: predictor or independent variable. REPEATED: a kind of input measured repeatedly across subjects (within-subjects). CONTROL: a kind of input to account for but not of main interest (covariate). OUTPUT: outcome or dependent variable.
<i>AddInputInteraction()</i>	Add an interaction/combination between input variables (INPUT, CONTROL, or REPEATED only) to your design.
<i>Test/Model Blocks (Shown in all tasks)</i>	
<i>GetTest()</i>	Select and run a statistical test based on the design you specified.
<i>GetModel()</i>	Build and train a predictive model based on the design you specified.

and the optional *astype()* variable blocks in the workspace.

The 15 test and model blocks covered the full range of analyses in the task sequence and were present in the same order on every task. Each block displayed its function call syntax with dropdown selectors for parameter values. An information tooltip [i] was available on each block, describing the function’s purpose and parameter semantics, pulled from the package documentation and truncated for brevity. For transparency all of the information tooltips provided to participants in the Control group are shown in Tables 4.4–4.7.

Ratings and Feedback Collection. Upon completing each task, either by submitting a correct answer or clicking “Give Up”, the dataset and workspace panels were hidden and a ratings sequence was presented. Anxiety and confidence were collected sequentially: the anxiety scale appeared first, and upon selection automatically advanced to the confidence scale after a brief 300ms delay to avoid double-clicks. Once both ratings were selected, a free-text feedback box appeared along with a “Next Task” button, which remained disabled until both ratings had been provided. A “Show Ratings” button also appeared if participants wished to change either of their initial ratings before advancing to the next task. After completing all 12

Table 4.4: Control interface block palette variable block information tooltip provided.

Block	Information Tooltip
<i>Variable Blocks (Shown in all tasks)</i>	
<i>astype()</i>	Converts a variable to a specified data type.
<i>Variable Blocks (Shown in tasks 11–12 only)</i>	
<i>StandardScaler()</i>	Standardize features by removing the mean and scaling to unit variance (mean=0 and standard deviation=1).
<i>OneHotEncoder()</i>	Encode categorical features as a one-hot numeric array. This creates a binary column for each category.
<i>OrdinalEncoder()</i>	Encode categorical features as an integer array. This results in a single column of integers (0 to n_categories - 1) per feature.
<i>LabelEncoder()</i>	Encode target labels with value between 0 and n_classes-1. This transformer should be used to encode target values, i.e. y, and not the input X.

tasks, participants were presented with a final free-text feedback box for overall comments about the study. Figure 4.4 shows the rating collection interface as presented to participants after task completion.

Interface Feedback/Hint Messages. Both interfaces used the same priority-ordered feedback system, designed to scaffold participants toward the correct answer without directly revealing it. Rather than displaying every applicable error at once, or defaulting to a generic “incorrect” message, the system was structured so that participants received the single most structurally fundamental piece of feedback first: intended to help them diagnose and correct their reasoning incrementally across submissions. When an answer was incorrect, the system evaluated the submission against a hierarchy of error conditions and displayed only the highest-priority applicable message, shown in Table 4.8.

Critically, this text feedback and the per-block color feedback operated as two independent channels. Each block added to the workspace turned red or green depending on whether any piece within it was incorrect, regardless of whether that error was the one described in the priority-ordered text. As a result, a participant could receive a text message describing one error (e.g., a missing test/model block) while a separate, unmentioned error was simultaneously visible via a red-highlighted variable block, allowing participants to potentially self-diagnose issues beyond the one explicitly named in the feedback text. The exception was the control interface’s *astype()* variable blocks, which were intentionally left gray regardless of correctness, since many of the parameters in the control test blocks overrode explicitly-declared types, making red/green feedback on these variable blocks potentially misleading.

Table 4.5: Control interface block palette penguin test/model block information tooltip provided.

Block	Information Tooltip
<i>Test/Model Blocks (Shown in all tasks)</i>	
<code>pg.corr()</code>	(Robust) correlation between two variables. x and y: first and second set of observations.
<code>pg.rm_corr()</code>	Repeated measures correlation. x and y: the two dependent variables. subject: the subject indicator.
<code>pg.chi2_independence()</code>	Chi-squared independence tests between two categorical variables. x and y: the variable names.
<code>pg.cochran()</code>	Cochran Q test. A special case of the Friedman test when the dependent variable is binary. dv: the dependent variable. within: the within-subject factor. subject: the subject/rater identifier.
<code>pg.anova()</code>	One-way ANOVA. dv: the dependent variable. between: the between-subject factor.
<code>pg.anova()</code>	Two-way ANOVA. dv: the dependent variable. between: list of the between-subject factors.
<code>pg.rm_anova()</code>	One-way repeated measures ANOVA. dv: the dependent variable. within: the within factor. subject: the subject identifier.
<code>pg.kruskal()</code>	Kruskal-Wallis H-test for independent samples. dv: the dependent variable. between: the between factor.
<code>pg.friedman()</code>	Friedman test for repeated measurements. dv: the dependent variable. within: the within-subject factor. subject: the subject/rater identifier.
<code>pg.mixed_anova()</code>	Mixed-design (split-plot) ANOVA. dv: the dependent variable. within: the within-subject factor (repeated measurements). between: the between-subject factor. subject: the between-subject identifier.

Implementation Notes. The control interface function call syntax for statsmodels and sklearn was simplified relative to the full library APIs for consistency with the penguin interface conventions. Specifically, *statsmodels* calls used the abbreviated *sm.manova()*, *sm.anova_lm.ols()*, and *sm.mixedlm()* forms rather

Table 4.6: Control interface block palette statsmodels test/model block information tooltip provided.

Block	Information Tooltip
<i>Test/Model Blocks (Shown in all tasks)</i>	
<code>sm.anova_lm.ols()</code>	Anova table for one or more fitted linear models. Create a Model from a formula and dataframe. formula: the formula specifying the model. The ~ separates outcome and predictors. The + adds variables. The * includes variables and their interaction.
<code>sm.mixedlm()</code>	Linear Mixed Effects Model. Create a Model from a formula and dataframe. formula: the formula specifying the model. The ~ separates outcome and predictors. The + adds variables. The * includes variables and their interaction.
<code>sm.manova()</code>	Multivariate Analysis of Variance. Create a Model from a formula and dataframe. formula: the formula specifying the model. The ~ separates outcome and predictors. The + adds variables. The * includes variables and their interaction.

Table 4.7: Control interface block palette scikitlearn test/model block information tooltip provided.

Block	Information Tooltip
<i>Test/Model Blocks (Shown in all tasks)</i>	
<code>skl.LinearRegression()</code>	Ordinary least squares Linear Regression. X: training data. y: target values.
<code>skl.LogisticRegression()</code>	Logistic Regression (aka logit, MaxEnt) classifier. X: training vector. y: target vector relative to X.

than the full module paths, and sklearn preprocessing functions were written as in-place column transformations (e.g., `df["x"] = OneHotEncoder(df["x"])`) rather than using *ColumnTransformer* pipelines. These simplifications preserved the conceptual decisions being evaluated, which test to select, which variables to assign to which parameters, and which preprocessing to apply, while removing boilerplate syntax unrelated to statistical understanding.

Hide Ratings

Correct! Please complete the following

Consider this definition: **Anxiety** is a feeling of fear or apprehension when presented with statistics content and problems or when gathering, processing and interpreting data.

Indicate how much **ANXIETY** you felt while working on this task.

No anxiety at all	A very small amount of anxiety	A small amount of anxiety	A moderate amount of anxiety	A large amount of anxiety	A very large amount of anxiety	An extreme amount of anxiety
☒	☐	☐	☐	☐	☐	☐

Consider this definition: **Confidence** is the sense of one's own ability to understand and complete actions required to accomplish specific statistics-related tasks.

Indicate how much **CONFIDENCE** you felt while working on this task.

No confidence at all	A very small amount of confidence	A small amount of confidence	A moderate amount of confidence	A large amount of confidence	A very large amount of confidence	An extreme amount of confidence
☐	☐	☐	☒	☐	☐	☐

Your feedback is very useful to us. Feel free to share any thoughts: What strategies did you use? What was confusing or helpful? Why did you rate your anxiety/confidence as you did?

Share your feedback, if you have any...

Next Task

Figure 4.4: The per-task ratings screen shown after task completion, with both anxiety and confidence ratings selected and the ratings panel re-opened for review. Anxiety is rated on a seven-point scale from “No anxiety at all” to “An extreme amount of anxiety”; confidence uses an identical scale. An open-ended feedback field is also provided.

4.3.9 Masking

Participants were not informed that the study involved two different interface groups; this ensured that their responses were not influenced by awareness of their group assignment. Group assignment used a blocked randomization procedure stratified by participants’ response to the gender demographic question. Within each gender stratum, assignment state persisted across the entire data collection period regardless of the time elapsed between participants: a participant was randomly assigned to either the experimental or control condition, the next participant from that same stratum was automatically assigned to the opposite condition, and the cycle then reset with a new random assignment. Researchers did not manually control group placement at any point. During analysis, group assignments and some demographic characteristics were included in statistical models, but no individual participant data was examined in a way that could introduce bias. This maintained a degree of partial blinding throughout the analysis process.

Table 4.8: Priority-ordered feedback message system. Upon each submission, the participant’s answer was evaluated against the hierarchy below and only the highest-priority applicable message displayed.

Priority	Condition	Message Displayed	Color
—	Starting feedback	<i>“Click “Check Answer” to check your progress.”</i>	Gray
1	Incomplete submission	<i>“Please fill out all dropdown fields before checking your answer.”</i>	Orange
2	Missing test or model	<i>“Please include a test or model.”</i>	Orange
3	Contradictory use	<i>“You used the same variable with different types or roles.”</i>	Orange
4	Missing variables	<i>“One or more required variables are missing.”</i>	Orange
5	Missing interaction	<i>“A required interaction between variables is missing.”</i>	Orange
6	Self-interaction	<i>“A variable cannot interact with itself.”</i>	Orange
7	Incorrect interaction	<i>“One or more variables in the interaction are incorrect.”</i>	Orange
8	Extra variables	<i>“You included extra or irrelevant variables or interactions.”</i>	Orange
9	Wrong test or model	<i>“The test or model does not match the task.”</i>	Orange
10	Incorrect variable types	<i>“One or more variable types are incorrect.”</i>	Orange
11	Incorrect variable roles	<i>“One or more variable roles are incorrect.”</i>	Orange
12	Missing preprocessing ^a	<i>“One or more variables need preprocessing (normalization or encoding).”</i>	Orange
—	Correct answer	<i>“Correct! Please complete the following.”</i>	Green
—	Give up	<i>“Please complete the following.”</i>	Gray

^a Preprocessing feedback applied to control Tasks 11–12 only.

4.3.10 Psychometrics

The correctness metrics introduced in this study are novel composite measures, and the affective scales are single-item instruments, neither of which lend themselves to traditional reliability estimation. Rather than applying inappropriate coefficients, scoring transparency is provided through the criteria tables below, and

construct validity is demonstrated through convergent evidence for the affective measures.

Table 4.9 reports the required minimum, mean, and maximum scored decisions per task and group. The required minimum is consistently higher for experimental participants, as variable types are always scored through variable block selection, whereas control participants typically only encounter type scoring when an *astype* block is explicitly applied. In practice, however, mean scored decisions were similar across groups on most tasks, as participants rarely triggered all optional scoring dimensions. Correctness scores are thus reasonably comparable in practice, though each group is evaluated on the reasoning decisions its interface elicits rather than on an identical set of criteria.

Table 4.9: Number of scored criteria by task and group

	Experimental				Control							
Task	Req.	Min	Obs.	Mean	Obs.	Max	Req.	Min	Obs.	Mean	Obs.	Max
1		7		7.7		11	5		6.7		10	
2		7		7.8		11	5		7.7		10	
3		7		7.7		11	5		6.4		10	
4		11		11.7		14	8		10.3		14	
5		9		9.1		12	7		8.2		10	
6		7		7.8		10	5		6.7		10	
7		13		13.0		13	10		11.3		14	
8		14		14.7		18	10		12.1		17	
9		14		14.7		15	10		12.2		18	
10		16		16.1		19	13		14.2		17	
11		16		16.8		20	16		17.4		23	
12		16		16.7		20	16		17.7		22	

Table 4.10 specifies the complete scoring criteria applied to each component and group. All scoring was automated and deterministic, rendering inter-rater reliability not applicable. Correctness scores represent the mean proportion of correct decisions across scored components for each participant, excluding inapplicable or un-scored decisions.

Convergent validity of the anxiety and confidence measures was assessed via repeated measures correlation across all tasks and participants. Consistent with prior literature [108, 134, 87, 66], anxiety and confidence were significantly negatively correlated ($r_{rm} = -.49$, $p < .001$, 95% CI $[-.53, -.45]$), supporting convergent validity. The moderate magnitude confirms the constructs are related but empirically distinguishable, supporting discriminant validity.

4.3.11 Conditions and Design

This study employed a between-subjects experimental design with random assignment to group. The independent variable, interface design, was actively manipulated by the researchers, with participants randomly

Table 4.10: Scoring criteria by component and group

Component	Scored	Experimental	Control
Test selection	Always	Declared via specific test block correct = 1 incorrect or absent = 0	Declared via specific test block correct = 1 incorrect or absent = 0
Variable inclusion	Always	Detected from presence within variable blocks; required & present = 1 required & absent = 0 extra & present = 0	Detected from presence within parameters of test block; required & present = 1 required & absent = 0 extra & present = 0
Variable role	Always	Declared via function name (e.g., <code>___Input</code> , <code>___Output</code>); correct = 1 incorrect or absent = 0	Declared via function parameter (e.g., <code>dv=</code> , <code>between=</code> , <code>within=</code>); correct = 1 incorrect or absent = 0
Variable type	Conditional	Always scored; declared via function name (e.g., <i>Categorical</i> ___, <i>Numerical</i> ___); correct or allowed = 1 incorrect or absent = 0	Scored only if <i>astype</i> explicitly applied, or <i>therapy_type</i> on Task 10; correct or allowed = 1 incorrect = 0
Preprocessing	Conditional	Never scored	Scored on Tasks 11–12 only, declared via encoding and scaling blocks; correct or allowed = 1 incorrect or absent = 0
Interaction	Conditional	Scored on Tasks 4,7,8,9,10 only, declared via <i>AddInputInteraction</i> block; correct pair = 1 incorrect, partial, or absent = 0	Scored on Tasks 4,7,8,9,10 only, declared via parameters in <i>anova2</i> / <i>mixed_anova</i> or from <i>*</i> operator in <i>anova_lm</i> / <i>mixedlm</i> / <i>manova</i> ; correct pair = 1 incorrect, partial, or absent = 0
ID variable	Conditional	Type never scored; inclusion and role scored on Tasks 5,7,10; on other tasks: inclusion = NaN, role scored if present	Type scored if <i>astype</i> used; inclusion and role scored on Tasks 5,7,10; on other tasks: inclusion = 0, role scored if present

assigned to either the Experimental or Control group based on their response to the gender demographic item to maintain balanced gender distributions across groups. Participants were not aware of the alternative group. The dependent variables, correctness, engagement, time, and self-reported anxiety and confidence, were measured repeatedly across 12 tasks, yielding a mixed design with one between-subjects factor (interface design: Experimental vs. Control) and one within-subjects factor (task: 1–12). The study was conducted online in a single session with no follow-up.

4.3.12 Data Diagnostics

Post-data-collection exclusions were applied prior to analysis as described in Section 4.3.1. Missing data arose naturally from task disengagement: participants who gave up on a task without any attempt produced no correctness or engagement data for that task. Rather than imputing missing values, disengagement was treated as a behavioral outcome and operationalized directly via the effort-weighted give-up rate. For inferential analyses requiring complete within-subjects data across all 12 tasks, participants with missing task responses were excluded rather than imputed, as the missing data pattern was not missing at random: disengagement was systematically related to task complexity and group.

Outliers were not removed from any analysis. Given the bounded nature of the primary measures (correctness: 0–1; anxiety and confidence: 1–7; give-up rate: 0–1), extreme values reflect behavioral variation rather than data entry errors or measurement artifacts.

Data distributions were examined prior to analysis. Correctness and give-up rate measures are bounded and non-normal by construction. Mixed ANOVAs were used as the primary inferential method, which are robust to moderate violations of normality given the sample sizes involved [43].

Data transformations. Effort-weighted engagement was computed as $-1/(1 + n)$ for tasks where the participant gave up, where n is the number of non-consecutive distinct attempts made before disengaging, and $1/n$ for successful task completions. This transformation preserves the ordinal structure of disengagement behavior while distinguishing both immediate give-up (-1.0) and efficient success (1.0) from high-effort disengagement and high-effort success (converging at 0). No other data transformations were applied.

4.3.13 Analytic Strategy

Primary hypotheses were evaluated using mixed ANOVAs with interface design (Experimental vs. Control) as the between-subjects factor and task (1–12) as the within-subjects factor. Greenhouse-Geisser corrections were applied where applicable. Effect sizes are reported as generalized eta squared (η_G^2). Pairwise comparisons were conducted using estimated marginal means with Tukey-Kramer correction where applicable.

Secondary and exploratory analyses, including component-level correctness breakdowns, role-specific trajectories, and subgroup comparisons by gender and disability status, are reported descriptively and interpreted with caution given their exploratory nature and, in some cases, small subgroup sizes. No correction for experiment-wise error was applied to exploratory analyses.

4.4 Results

4.4.1 Engagement

4.4.1.1 Giving Up and Weighted Engagement

The raw give-up rate across tasks was substantially lower in the experimental group than in the control group (25.9% vs. 53.3%), indicating that control participants abandoned tasks at roughly twice the rate of experimental participants. However, raw give-up rate treats all task abandonment equivalently regardless of how much effort preceded them. Among give-ups in the control group, 90 of 351 (25.6% of give-up tasks, 13.7% of all tasks) occurred without any attempt, compared to 46 of 182 (25.3% of give-up tasks, 6.5% of all tasks) in the experimental group, a similar proportion within give-ups across groups, suggesting that immediate disengagement was comparably distributed but disengagement was nearly twice as common in control.

To capture the full spectrum of engagement behavior, from immediate abandonment through effortful persistence to efficient success, we computed an *effort-weighted engagement* metric applied symmetrically to both give-ups and successes. A mixed ANOVA revealed significant main effects of design and task, as well as a significant interaction (Table 4.11), indicating that engagement trajectories diverged across groups as task complexity increased.

Table 4.11: Mixed ANOVA for effort-weighted engagement

Effect	df	dfe	MSE	F	η_G^2	p
Design	1	110	1.71	7.08	.033	.009
Task	7.27	800.02	0.21	19.13	.076	< .001
Design $\times$ Task	7.27	800.02	0.21	23.59	.092	< .001

Post hoc comparisons revealed that on Task 1 (Correlation), control participants showed significantly higher weighted engagement than experimental participants ($t = 7.47$, $p < .001$), possibly slightly inflated by ordering of the test palette placing correlation first and enabling rapid selection without reasoning, mentioned in qualitative feedback. From Task 2 onward, this advantage reversed, with experimental participants showing equal or higher weighted engagement on all remaining tasks. Differences were non-significant on tasks 3–7 but became large and highly significant on Tasks 8–12, where control participants showed sharply declining engagement while experimental participants remained relatively stable (Table 4.12).

Illustrated in Figure 4.5, the mean weighted engagement score for experimental participants remained above the persisted engagement threshold across all 12 tasks, indicating sustained net positive engagement even on the most complex statistical and machine learning tasks. By contrast, control participants crossed below this threshold from Task 8 onward, reflecting collective disengagement as task complexity increased. The distribution of weighted engagement scores (shown in the violin plots) further reveals that experimental

Table 4.12: Post-hoc comparisons for effort-weighted engagement by task (Control – Experimental)

Task	Ctrl <i>M</i> (SD)	Exp. <i>M</i> (SD)	Estimate	SE	<i>t</i> (110)	<i>p</i>
1 (Correlation)	0.609 (0.461)	0.074 (0.281)	0.535	0.072	7.472	< .001***
2 (1-Way ANOVA)	0.048 (0.416)	0.351 (0.503)	-0.303	0.088	-3.459	< .001***
3 (Chi-Squared)	0.413 (0.620)	0.341 (0.556)	0.072	0.111	0.650	.517
4 (2-Way ANOVA)	0.142 (0.533)	0.315 (0.531)	-0.173	0.101	-1.723	.088
5 (RM ANOVA)	0.073 (0.480)	0.080 (0.421)	-0.007	0.085	-0.084	.933
6 (Kruskal-Wallis)	0.119 (0.557)	0.250 (0.456)	-0.131	0.096	-1.368	.174
7 (Mixed ANOVA)	0.165 (0.593)	0.113 (0.435)	0.051	0.098	0.523	.602
8 (MANOVA)	-0.075 (0.692)	0.304 (0.601)	-0.379	0.122	-3.102	.002**
9 (2-Way ANCOVA)	-0.198 (0.569)	0.134 (0.522)	-0.332	0.103	-3.221	.002**
10 (Mixed ANCOVA)	-0.191 (0.525)	0.004 (0.466)	-0.195	0.094	-2.086	.039*
11 (Linear Reg.)	-0.421 (0.482)	0.252 (0.528)	-0.673	0.096	-7.027	< .001***
12 (Logistic Reg.)	-0.420 (0.492)	0.324 (0.627)	-0.744	0.107	-6.956	< .001***

participants consistently retained a higher proportion of scores concentrated in the upper ‘success’ zone throughout, while control participants showed a progressively greater proportion of scores concentrated in the lower ‘give-up’ zone as tasks advanced.

4.4.1.2 Frustration

Frustration, operationalized as repeated identical submissions made without modification, was more prevalent in the control group than in the experimental group, though observed in both groups. Consistent with prior HCI research identifying repetitive interaction behaviors as a reliable behavioral signal of frustration [122, 116, 65], frustration bouts were identified as either two or more identical submissions immediately followed by task abandonment, or three or more consecutive identical submissions during task completion.

In the control group, 45 of 59 participants (76.3%) exhibited at least one frustration bout, compared to 34 of 61 participants (55.7%) in the experimental group. Among participants who exhibited frustration, the average interval between repeated identical submissions was 8.14 seconds in the control group and 5.35 seconds in the experimental group, suggesting that control participants took longer between identical submissions, possibly reflecting reconsideration, while experimental participants repeated submissions more rapidly when frustrated.

Task abandonment following frustration was substantially more common in the control group. Of 659 completed control-group tasks, 352 (53.4%) resulted in give-up, and 52 of these (14.8% of give-ups; 7.9% of all tasks) ended immediately following a frustration bout with no further modification. In contrast, of 703 completed experimental-group tasks, 183 (26.0%) resulted in give-up, and only 21 (11.5% of give-ups; 3.0%

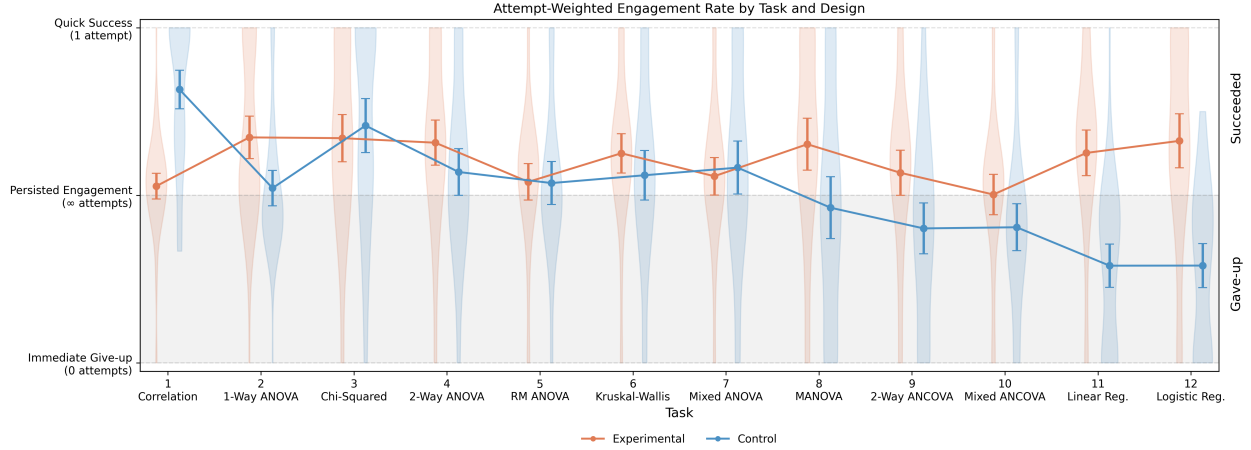


Figure 4.5: Effort-weighted engagement by task and group. The upper portion of the chart reflects successful task completion, the lower portion reflects task abandonment. Violin plots show the full distribution; points and error bars represent group means with 95% confidence intervals.

of all tasks) ended immediately after a frustration bout. In many experimental-group cases, participants recovered from frustration bouts and continued modifying their responses rather than abandoning the task.

Of the 8 study non-completers, two control-group participants discontinued the study immediately following frustration bouts. One exited during Task 5 after 87 attempts, 60 of which were frustration clicks averaging 1.60 seconds between repeated submissions; the other exited after a brief frustration bout in Task 4. Both had exhibited frustration in earlier tasks despite completing the practice task. One experimental participant exhibited a single frustration bout on Task 1, and was also the only non-completer to have skipped the majority of the practice task. Despite this, the participant successfully completed Tasks 2–4 without further frustration, each within a few attempts, before discontinuing for unknown reasons, suggesting the frustration in Task 1 may have simply reflected incomplete interface training.

4.4.2 Performance

4.4.2.1 Correctness

Eventual correctness, whether a participant ultimately reached a correct solution, was 74.0% for experimental and 46.6% for control, the direct complement of the raw give-up rates reported above. This was not adopted as a primary outcome measure. Because control participants gave up at substantially higher rates, eventual correctness would systematically undercount control performance for reasons unrelated to reasoning quality, conflating disengagement with analytical ability. Thus, two pre-feedback metrics were instead used to isolate reasoning performance: *first-attempt correctness*, capturing the accuracy of the full workspace at the first submission, and *first-intent correctness*, capturing the accuracy of each reasoning decision the first time it was made, accumulated across attempts until all elements had been evaluated, unless preempted by a

give-up.

Both metrics were analyzed using mixed ANOVAs with interface design as the between-subjects factor and task as the within-subjects factor. Participants who never submitted an attempt on a given task received a correctness score of zero for that task, as excluding them would have disproportionately removed control participants and introduced substantial group imbalance. Only those who completed all tasks were included ($N = 112$).

Table 4.13: Mixed ANOVAs for first-attempt and first-intent correctness

Effect	First-Attempt			First-Intent		
	<i>F</i>	η_G^2	<i>p</i>	<i>F</i>	η_G^2	<i>p</i>
Design	5.36	.022	.022	27.80	.113	< .001
Task	6.62	.031	< .001	7.81	.034	< .001
Design $\times$ Task	8.23	.039	< .001	10.74	.046	< .001

Note. Greenhouse-Geisser correction applied. $N = 112$ completers.

Overall, experimental participants outperformed control participants on both metrics. For first-attempt correctness, experimental participants scored higher ($M = 0.635$, $SD = 0.352$) than control participants ($M = 0.539$, $SD = 0.337$). For first-intent correctness, the gap was larger: experimental ($M = 0.781$, $SD = 0.265$) versus control ($M = 0.582$, $SD = 0.318$). Both analyses revealed significant main effects of design and task, and significant interactions (Table 4.13), indicating that correctness differed between groups and that this difference grew with task complexity. The larger design effect for first-intent correctness ($\eta_G^2 = .113$) compared to first-attempt correctness ($\eta_G^2 = .022$) suggests that first intent more precisely captures reasoning quality by isolating each decision from the noise of iterative submission behavior, consistent with the expectation that experimental participants' incremental building style would inflate attempt counts even for correct reasoning.

Post hoc comparisons revealed that experimental participants outperformed control on nearly all tasks for both metrics (Tables 4.14 and 4.15). For first-attempt correctness, control outperformed experimental on Task 1 ($t = 3.28$, $p = .001$), with no significant differences on Tasks 2–3 and experimental advantages emerging consistently from Task 4 onwards. For first-intent correctness, control outperformed experimental on Task 1 ($t = 2.43$, $p = .017$), with no significant difference on Task 3, and experimental advantages emerging from Task 2 onwards. The patterns on Tasks 1 and 3 are discussed further below.

4.4.2.2 Learning Trajectories

A consistent theme across the correctness data is that the experimental interface imposed an initial adjustment cost that gave way to improved performance and conceptual development over time, while the control interface supported early performance through familiarity but provided diminishing returns as task

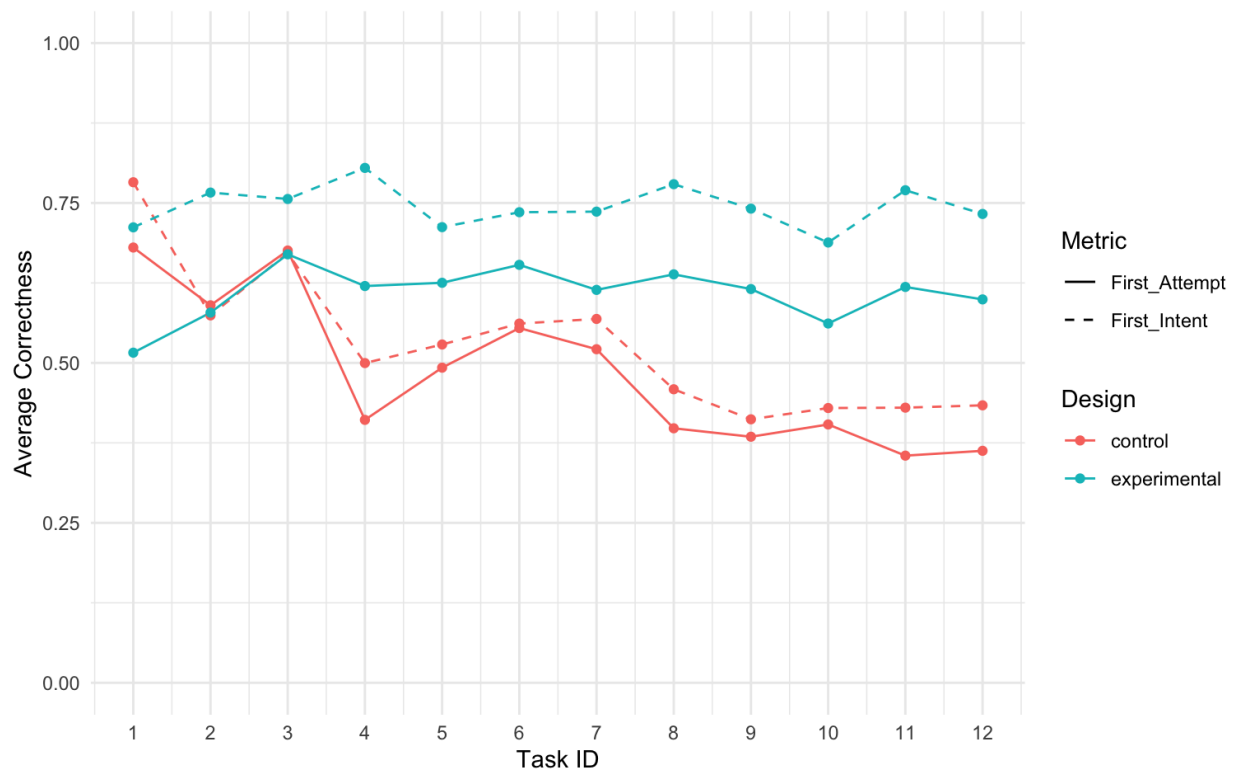


Figure 4.6: First-attempt and first-intent correctness by task (give-ups scored as incorrect).

Table 4.14: Post-hoc comparisons for first-attempt correctness by task (Control – Experimental)

Task	Ctrl <i>M</i> (SD)	Exp <i>M</i> (SD)	Estimate	SE	<i>t</i> (110)	<i>p</i>
1 (Correlation)	0.745 (0.337)	0.536 (0.338)	0.209	0.064	3.277	.001**
2 (1-Way ANOVA)	0.641 (0.272)	0.604 (0.381)	0.037	0.063	0.583	.561
3 (Chi-Squared)	0.740 (0.323)	0.700 (0.324)	0.040	0.061	0.654	.514
4 (2-Way ANOVA)	0.468 (0.299)	0.651 (0.372)	-0.183	0.064	-2.863	.005**
5 (RM ANOVA)	0.541 (0.263)	0.645 (0.288)	-0.104	0.052	-1.991	.049*
6 (Kruskal-Wallis)	0.611 (0.324)	0.685 (0.311)	-0.074	0.060	-1.231	.221
7 (Mixed ANOVA)	0.579 (0.325)	0.646 (0.338)	-0.067	0.063	-1.070	.287
8 (MANOVA)	0.443 (0.365)	0.657 (0.386)	-0.214	0.071	-3.007	.003**
9 (2-Way ANCOVA)	0.430 (0.306)	0.647 (0.349)	-0.218	0.062	-3.504	< .001***
10 (Mixed ANCOVA)	0.450 (0.278)	0.588 (0.326)	-0.138	0.058	-2.406	.018*
11 (Linear Reg.)	0.393 (0.329)	0.648 (0.381)	-0.255	0.068	-3.773	< .001***
12 (Logistic Reg.)	0.396 (0.360)	0.629 (0.409)	-0.233	0.073	-3.197	.002**

complexity increased.

This pattern is visible from Task 1. Control participants outperformed experimental participants significantly on both first-attempt ($t = 3.49$, $p < .001$) and first-intent correctness ($t = 4.04$, $p < .001$), reflecting two advantages that did not persist: correlation is the most familiar statistical test in the study, and it appeared first in the control palette due to ordering. Experimental participants, encountering a novel block-based interface for the first time with no test names, reported needing to explore and orient before understanding what was expected of them. Several participants noted during debriefing that Task 1 served as a learning experience for the interface itself. From Task 2 and on, this initial disadvantage reversed and experimental participants began to outperform control on nearly every subsequent task. Control participants showed a noticeable jump in correctness on Task 3 (Chi-Squared) where they performed similarly to the experimental group. Qualitative feedback from several control participants revealed this was due to keyword matching: *chi2_independence* contains the word “independence,” which directly mapped to the task description. This pattern inadvertently supports the descriptive naming findings of the first study and previous work regarding “say what you do” naming conventions.

The correctness gap narrowed on Task 6 (Kruskal-Wallis), driven by type reasoning for the *satisfaction* variable. This variable required an *ordered* type declaration in the experimental group, but 34 of 58 experimental participants declared it *numeric*. Control participants were not penalized for this, as *numeric* is an accepted encoding for *kruskal* test, resulting in control outperforming experimental on type correctness for this task (63.6% vs. 48.7%). However, this appeared to have a learning impact for the experimental group: on Task 10 (Mixed ANCOVA), experimental participants substantially outperformed control when correctly

Table 4.15: Post-hoc comparisons for first-intent correctness by task (Control – Experimental)

Task	Ctrl <i>M</i> (SD)	Exp <i>M</i> (SD)	Estimate	SE	<i>t</i> (110)	<i>p</i>
1 (Correlation)	0.836 (0.238)	0.733 (0.212)	0.103	0.043	2.430	.017*
2 (1-Way ANOVA)	0.628 (0.186)	0.802 (0.241)	-0.173	0.041	-4.234	< .001***
3 (Chi-Squared)	0.736 (0.311)	0.791 (0.247)	-0.055	0.053	-1.047	.297
4 (2-Way ANOVA)	0.554 (0.279)	0.845 (0.273)	-0.291	0.052	-5.583	< .001***
5 (RM ANOVA)	0.578 (0.245)	0.748 (0.210)	-0.170	0.043	-3.943	< .001***
6 (Kruskal-Wallis)	0.615 (0.299)	0.774 (0.223)	-0.159	0.050	-3.208	.002**
7 (Mixed ANOVA)	0.625 (0.309)	0.775 (0.268)	-0.150	0.055	-2.740	.007**
8 (MANOVA)	0.517 (0.331)	0.820 (0.289)	-0.303	0.059	-5.165	< .001***
9 (2-Way ANCOVA)	0.452 (0.304)	0.779 (0.297)	-0.327	0.057	-5.765	< .001***
10 (Mixed ANCOVA)	0.478 (0.277)	0.724 (0.284)	-0.246	0.053	-4.629	< .001***
11 (Linear Reg.)	0.476 (0.365)	0.810 (0.290)	-0.334	0.062	-5.375	< .001***
12 (Logistic Reg.)	0.473 (0.399)	0.771 (0.327)	-0.297	0.069	-4.322	< .001***

identifying *therapy_type* as categorical despite its numeric default encoding in 51.7% of cases compared to 13.0% for control.

Component-level trajectories for the two most conceptually demanding role types further illustrate this distinction. For the repeated-measures role, experimental participants showed a gradual upward trajectory across Tasks 5, 7, and 10 (13.6%, 24.1%, 32.8%) as the repeated role became more reliably understood, while control participants rose then fell (30.9%, 40.7%, 20.4%), a decline coinciding with the syntax shift from *within=* parameter notation to formula syntax on Task 10, suggesting their performance had been anchored to a parameter label, constrained syntax and information provided rather than the underlying concept. For the covariate role, experimental participants identified it at roughly twice the rate of control participants on both Tasks 9 and 10 (24.1% vs. 11.1%; 44.8% vs. 22.2%), despite both groups improving across the two tasks. Across both role types, explicitly declaring the role appears to have produced more durable conceptual encoding than assigning variables to parameter slots whose names carried the structural meaning implicitly.

4.4.3 Affective Outcomes

4.4.3.1 Construct Reliability

Convergent and discriminant validity of the anxiety and confidence measures were assessed using repeated measures correlation, which partials out between-person variance to estimate within-person associations across the 12 tasks ($N = 112$, $df = 1231$). Anxiety and confidence were significantly negatively correlated ($r_{rm} = -.49$, $p < .001$), confirming convergent validity, consistent with prior work showing that lower

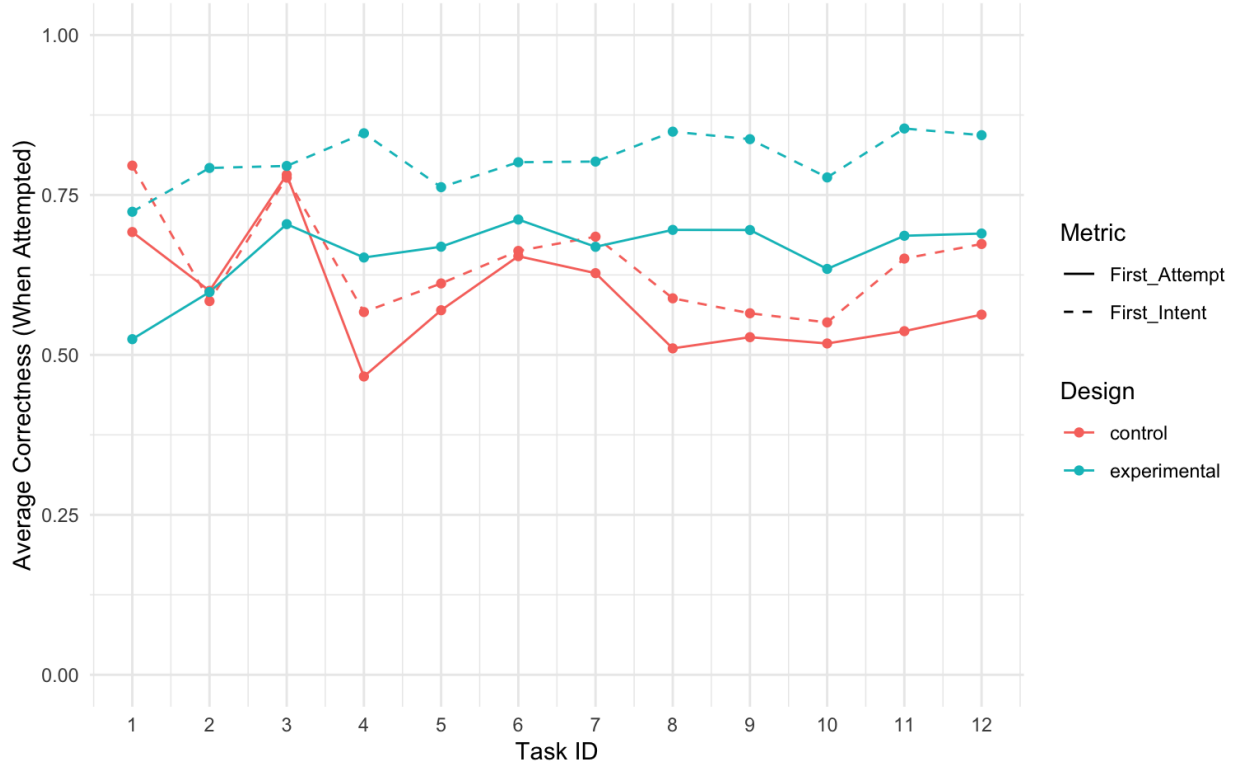


Figure 4.7: Correctness by task, conditional on attempting the task.

statistics anxiety is associated with higher self-efficacy/confidence profiles among students [66]: the two scales moved in opposite directions within participants across tasks, as expected for measures of opposing affective states.

4.4.3.2 Anxiety

A mixed ANOVA on anxiety ratings from the 112 completers revealed no significant main effect of design ($F(1, 110) = 2.51$, $\eta_G^2 = .014$, $p = .116$), but significant main effects of task ($F(5.95, 654.23) = 3.21$, $\eta_G^2 = .012$, $p = .004$) and a significant design–task interaction ($F(5.95, 654.23) = 4.36$, $\eta_G^2 = .016$, $p < .001$), indicating that anxiety trajectories diverged across groups as task complexity increased (Table 4.16). Post hoc Tukey–Kramer comparisons showed control participants reported significantly higher anxiety than experimental participants on Tasks 2, 11, and 12, with non-significant trends in the same direction on Tasks 4–10.

Interpreting these results requires careful examination of how anxiety ratings related to task outcome. Task-level ratings were collected immediately after each task concluded, regardless of whether the participant succeeded or gave up, meaning that ratings following disengagement reflect the affective state produced included the disengagement itself rather than the task experience alone. To examine this, task-level responses were classified into four outcome states: *Disengaged* (gave up without adding any blocks to the workspace),

Table 4.16: Mixed ANOVA for anxiety ratings

Effect	df	MSE	F	η_G^2	p
Design	1, 110	17.69	2.51	.014	.116
Task	5.95, 654.23	1.98	3.21	.012	.004
Design $\times$ Task	5.95, 654.23	1.98	4.36	.016	< .001

Note. Greenhouse-Geisser correction applied. $N = 112$ completers.

Explored (a small subset who gave up without submitting an answer but added at least one block to the workspace, thus not completely disengaged, $n_{\text{ctrl}} = 14$, $n_{\text{exp}} = 16$), *Attempted Give-up* (gave up after at least one submission), and *Succeeded* (reached the correct solution after one or more attempts). Mean anxiety and confidence ratings for each state are shown in Table 4.17.

Table 4.17: Mean anxiety and confidence by group and task outcome

		Gave-Up			Succeeded
		Disengaged	Explored	Attempted	
Control	Anxiety	2.83	3.50	3.34	2.85
	Confidence	1.37	1.50	1.97	3.41
	n	76	14	261	308
Experimental	Anxiety	4.10	2.69	3.94	2.30
	Confidence	2.77	3.81	2.37	4.21
	n	30	16	136	521

Note. Ratings on a 7-point scale. Disengaged = gave up without adding blocks to the workspace.

Explored = gave up after adding blocks but without submitting. Attempted = gave up after at least one submission. Explored cells are small ($n = 14$, $n = 16$); interpret cautiously.

The pattern in Table 4.17 reveals that the non-significant main effect of design on anxiety is substantially attributable to how disengagement affected affective ratings differently across groups. In the control group, participants who disengaged without touching the workspace reported mean anxiety of 2.83, virtually identical to the mean anxiety of control participants who succeeded ($M = 2.85$). However, confidence tells an entirely different story: disengaged control participants reported mean confidence of 1.37, compared to 3.41 for those who succeeded, a difference of more than two points on a 7-point scale. The same low anxiety rating therefore describes two qualitatively different affective states. Among control participants who disengaged without attempting, 44.7% reported simultaneous 1/1 ratings and 75.0% rated confidence at the scale floor. Task-level feedback from these participants suggests these floor ratings were intentional rather than incidental: “I honestly have no idea what anything means so I don’t feel anxiety or pressure to answer”;

“I have no confidence anymore and don’t want to answer”; “again no idea so no anxiety or confidence”; “I knew I was not able to get it.” Critically, these comments appeared on Tasks 4–11, after the practice task and several completed tasks, suggesting the pattern was not an artifact of unfamiliarity with the interface but a response to the statistical content itself, consistent with a state of low engagement following repeated unresolved failure rather than genuine ease [31, 109].

The experimental group presents a different picture. Experimental participants who disengaged without touching the workspace reported the highest anxiety of any group ($M = 4.10$), with confidence remaining moderate ($M = 2.77$) and only 6.7% giving 1/1 ratings. Among the 8 experimental participants who gave 1/1 ratings, the two who left feedback did so on Tasks 2 and 3 (“I am completely confused while looking at these questions”; “I don’t know what language this is”), consistent with early interface orientation rather than the pattern of repeated statistical failure observed in the control group. Taken together with the elevated anxiety and moderate confidence at the group level, this suggests that experimental participants who disengaged gave responses that may be more consistent with active confusion or frustration than with the low-anxiety, low-confidence disengagement seen in the control group.

Taken together, these patterns suggest that low anxiety ratings following disengagement in the control group reflect a substantive affective state rather than a measurement artifact. While the spatial proximity of the give-up button to the lowest response option may have facilitated floor-level selections [17, 98], the qualitative responses above suggest the low ratings were deliberate rather than incidental. This state was more prevalent in the control group because the control interface produced the psychological conditions for it more frequently. The implications for this are discussed in Section 4.6.4.

4.4.3.3 Confidence

Confidence showed a large and highly significant main effect of design ($F(1, 110) = 36.74$, $\eta_G^2 = .152$, $p < .001$), a significant main effect of task ($F(7.58, 834.26) = 2.44$, $\eta_G^2 = .010$, $p = .015$), and a significant design–task interaction ($F(7.58, 834.26) = 10.63$, $\eta_G^2 = .043$, $p < .001$), indicating that confidence diverged substantially between groups and that this divergence grew with task complexity (Table 4.18). Unlike anxiety, the floor effect that attenuated anxiety differences did not suppress confidence differences to the same degree: the design effect for confidence ($\eta_G^2 = .152$) is an order of magnitude larger than for anxiety ($\eta_G^2 = .014$), and the pattern in Table 4.17 may give a clue as to why. Confidence remained near the floor across all give-up outcomes in the control group regardless of how much effort preceded disengagement, while experimental participants maintained substantially higher confidence even when they gave up. Therefore, confidence may be the more sensitive measure of the between-group affective difference in this study, both because it was less susceptible to floor suppression following disengagement and because it tracked the cumulative experience of the interface across tasks more precisely than anxiety did.

Post hoc Tukey-Kramer comparisons revealed no significant difference on Task 1, consistent with the interface orientation period observed for experimental participants, but significant experimental advantages

Table 4.18: Mixed ANOVA for confidence ratings

Effect	df	MSE	F	η_G^2	p
Design	1, 110	13.66	36.74	.152	< .001
Task	7.58, 834.26	1.56	2.44	.010	.015
Design $\times$ Task	7.58, 834.26	1.56	10.63	.043	< .001

Note. Greenhouse-Geisser correction applied. $N = 112$ completers.

on every subsequent task (Table 4.19). The largest differences emerged on Tasks 11 and 12, where experimental participants reported confidence nearly two points higher than control participants on the 7-point scale. Experimental confidence grew significantly across tasks while control confidence decreased, mirroring the diverging correctness and engagement trajectories reported above.

Table 4.19: Post-hoc comparisons for confidence by task (Control – Experimental)

Task	Ctrl M (SD)	Exp M (SD)	Estimate	SE	$t(110)$	p
1 (Correlation)	3.19 (1.39)	3.12 (1.17)	0.065	0.242	0.266	.790
2 (1-Way ANOVA)	2.54 (1.41)	3.59 (1.56)	-1.049	0.281	-3.729	< .001***
3 (Chi-Squared)	3.17 (1.60)	3.84 (1.48)	-0.678	0.291	-2.328	.022*
4 (2-Way ANOVA)	2.74 (1.26)	3.95 (1.42)	-1.208	0.254	-4.745	< .001***
5 (RM ANOVA)	2.57 (1.42)	3.34 (1.28)	-0.771	0.255	-3.020	.003**
6 (Kruskal-Wallis)	2.83 (1.48)	3.72 (1.56)	-0.891	0.288	-3.093	.003**
7 (Mixed ANOVA)	2.56 (1.31)	3.72 (1.48)	-1.169	0.266	-4.401	< .001***
8 (MANOVA)	2.26 (1.14)	3.97 (1.57)	-1.706	0.260	-6.557	< .001***
9 (2-Way ANCOVA)	2.31 (1.36)	4.07 (1.66)	-1.754	0.288	-6.088	< .001***
10 (Mixed ANCOVA)	2.37 (1.46)	3.60 (1.53)	-1.233	0.283	-4.356	< .001***
11 (Linear Reg.)	2.00 (1.39)	4.16 (1.72)	-2.155	0.297	-7.253	< .001***
12 (Logistic Reg.)	2.17 (1.48)	4.29 (1.63)	-2.126	0.295	-7.211	< .001***

4.4.4 Exploratory Analyses

Gender No significant gender differences were observed in correctness, anxiety, confidence, give-up rate, or engagement within either group, with the exception of a trending correctness advantage for female participants in the control group ($p = .065$) that did not replicate in the experimental group. Mean anxiety was identical across genders in both groups (control: $M = 3.05$ for both; experimental: $M = 2.70$ for both), while confidence was slightly higher for experimental participants regardless of gender (female: $M = 3.57$ vs. $M = 2.72$; male: $M = 3.86$ vs. $M = 2.54$), consistent with the overall group-level patterns.

The general absence of gender effects should be interpreted in light of the sample composition. The sample was male-dominated (79 male, 31 female), limiting statistical power for gender comparisons. This is notable given that Trassi et al. (2022) found that the large majority of statistical anxiety research has been conducted with predominantly female samples drawn from human science and health fields, with only 15% of studies drawing from STEM fields [134]; our male-dominated STEM sample is thus not well represented in the existing literature base, which may partly explain why gender effects here diverge from the more commonly reported pattern of women reporting higher statistical anxiety [11, 47, 37, 20, 90]. The absence of a gender difference also aligns with Mji (2009), who similarly found no significant gender gap among technology students [95]. A gender-balanced replication of this study outside STEM fields, may reveal whether the experimental interface supports genders differently based on domain.

Disability Status An exploratory analysis examined whether outcomes differed by disability status within each group ($n = 20$ with disability: 8 experimental, 12 control). Given the small subsample, all findings should be interpreted as preliminary and hypothesis-generating rather than confirmatory.

Within the control group, participants with disabilities reported higher anxiety ($M = 3.74$, $SD = 1.56$) than those without ($M = 2.87$, $SD = 1.03$), though this difference did not reach significance, $t(13.89) = -1.82$, $p = .091$. This pattern did not emerge in the experimental group ($M_{\text{disability}} = 2.93$, $SD = 1.44$; $M_{\text{no disability}} = 2.66$, $SD = 1.19$), $t(8.60) = -0.49$, $p = .634$. Comparing across groups among participants with disabilities only, experimental participants reported significantly higher confidence ($M = 3.39$, $SD = 0.85$) than control participants ($M = 2.47$, $SD = 0.95$), $t(16.31) = -2.25$, $p = .039$, and a trending reduction in give-up rate ($M = 24.0\%$, $SD = 36.0\%$ vs. $M = 52.8\%$, $SD = 32.8\%$), $t(14.14) = 1.82$, $p = .091$, consistent with the overall group-level patterns.

Together, these findings tentatively suggest that the experimental interface’s explicit structure may have particularly benefited participants with disabilities, reflected in significantly higher confidence and a trending reduction in task abandonment. The anxiety pattern is less clear: a trending disability-related gap appeared within the control group but not the experimental group, yet the direct cross-group comparison was not significant. Due to such small sub groups, replication is needed before drawing firm conclusions.

4.5 Qualitative Feedback

Participants were invited to leave optional open-ended comments after each of the 12 tasks and again at the end of the study. Across the twelve per-task prompts, 34 of 61 experimental participants (56%) and 35 of 59 control participants (59%) provided at least one comment, yielding 230 and 239 total per-task entries respectively. End-of-study comments were provided by 22 participants in each group (36% experimental, 37% control). Feedback is used here interpretively rather than as a formal thematic analysis [15]: task-level comments are examined selectively to contextualize specific quantitative patterns, and end-of-study comments are organized around patterns that recurred across the full sample.

4.5.1 Task-Level Feedback

Two diverging arcs. Task-level comments reveal two qualitatively different trajectories across the study. In the experimental group, early comments reflect genuine disorientation with an unfamiliar interface, but this resolved rapidly and consistently. One participant’s sequence captures the arc directly: “Well it was stressful because the top problem was giving me issues” (Task 1); “am I just completely not doing it correctly, maybe it’s just on my end” (Task 2); “ok so I got both parts correct. What is going on” (Task 3); “Ok so now I understand what’s going on, I feel like a silly goose now” (Task 4). Several others described the same transition more briefly: “code blocks scare me” (Task 1) to “okay this isn’t as hard as I thought it was” (Task 3); “Wording of what each function does is hard” (Task 1) to “I understand it more” (Task 2). The pattern is consistent: initial confusion, quick reorientation, and sustained engagement thereafter. By the later tasks, experimental participants described the interface in terms of logic and strategy: “the more I understood the program, the more intuitive and obvious the answers became,” “now that I’m used to this. No crazy strategies needed just follow what makes sense logically,” and “I figured out a strategy... and everything seemed natural and easy.”

In the control group the arc ran in the opposite direction. Early comments often reflected confidence built on familiarity of statistical concepts, function names and ordering of the block palette: “the task was quite easy,” “I just thought to myself, what variables do I need,” “It seemed almost too simple,” “I have a pretty good understanding of the use of correlation in statistics,” “I just picked the first one,” and “I just went with my gut and used the first option, which was the basic correlation.” As complexity grew, however, these surface-level strategies stopped working. One participant’s comments are representative: Task 1, “I had no idea I was going to get it on the first try”; Tasks 4–10, “I do not know what the rest of the functions in the task section do” (repeated verbatim across seven consecutive tasks); with an additional comment on Tasks 9–10, “The tasks keep getting more complex. I don’t even attempt at this point.” Where experimental participants described learning the interface, control participants more often described reaching its ceiling.

The difference between groups was not one of effort but of what the interface required participants to think about. Experimental participants described reasoning about variable meaning: whether satisfaction was ordered or categorical, whether lighting condition was repeated or input, whether gender served as a control variable. Control participants described reasoning about the interface itself: counting arguments (“I simply saw how many columns there were and deduced that the function would have to have the same amount of variables”, a strategy mentioned by a large portion of the control participants), systematic enumeration (“I’m just iterating. If it has 3 variables, look for a function with 3 arguments till correct test is found... then shuffle placement of arguments till correct”), and description scanning (“I always refer to the info buttons to know which block to use. They are very helpful”). These strategies were effective on simpler tasks and broke down as complexity grew: “My previous strategies do not work at all” (Task 7); “Tried iterating but could not keep focus to do counting correctly” (Task 10); and “I saw the different variable types and knew there was no shot” (Task 11).

Keyword matching on Task 3. The near-parity in correctness on Task 3 is explained by a pattern in control participants’ comments: several independently identified the chi-squared test by matching the word “independence” in the task description to *pg.chi2_independence*, with several participants noting this directly: “the problem mentioned independence and one of the models had the same word”; “I remembered that there was a test labeled ‘independence’”; “I saw that the chi-squared function described independence.” These reflect surface-level keyword matching rather than structural reasoning.

4.5.2 End-of-Study Feedback

A learning curve giving way to intuition and enjoyment. Experimental participants overwhelmingly described an early period of disorientation followed by a shift toward fluency: “The study started off very shaky but as time went on I started troubleshooting the problems and it made a lot of sense”; “most questions built upon each other or introduced one new thing so it did not make it too overwhelming to learn”; “Fixing the errors was initially challenging, but once I fully understood the task, the process became straightforward once again. Overall, it was a great study and a great experience.” Many described the shift in terms of enjoyment and intrinsic motivation: “I enjoyed figuring it out and trying new methods (almost like a puzzle),” “it was fun learning how to do statistics,” “Although I wished I got a clearer understanding for some parameters (mainly Repeated and Ordered)... I managed to understand what they did and know when to use them... I found myself more confident with my skills”; “This study was an enjoyable experience”; “I liked the problem-solving challenge and the challenge of interpreting the descriptions”; “I feel like I learned more about how to do data science effectively using these exercises.” A subset of the control group gave neutral or positive feedback: “cool study. I got nervous”; “The study was fine itself”; “I think the premise was good however my lack of knowledge leaves my ability to properly respond at a level I find underwhelming”; “this was a learning experience but I got more confident as I went along”; “I liked this study! It was much less stressful than I thought mainly because the functions were already implemented.” However, the framing of tasks as puzzles rather than tests, was essentially absent from control end-of-study comments, which more often described frustration, lack of knowledge, or confusion: “The first few I kind of got, but after the 5th or so task I couldn’t really get anything running,” “I clearly don’t yet have the required knowledge,” “I got significantly more annoyed as time went on, descriptions of functions and variables do not help,” and “Honestly going in I thought I would do better than I did.” Where experimental participants typically described learning how to use the interface despite feeling initially unsure, control participants often described learning that they were limited by their current knowledge and the interface did not bridge this gap.

Effort exhaustion and disengagement in the control group. A subset of control participants described a qualitative shift from engagement to apathy that the give-up rate captures only partially: “after a while apathy took over and I didn’t have the energy to keep trying”; “over time, learned helplessness instills

and it’s hard to overcome it”; “it was frustrating to not be able to understand why my solutions didn’t work.” These accounts suggest that control participants who gave up were frequently not participants who made no effort, but participants whose effort had produced no intelligible progress. Experimental participants rarely described this kind of exhaustion. Where frustration appeared it was more often localized: “I don’t know how those two are supposed to interact but I was getting frustrated and decided to move on and come back later”, suggesting frustration was real, but not often terminal in the experimental group.

4.6 Discussion

4.6.1 Support of Original Hypotheses

Engagement. Experimental participants gave up on 25.9% of tasks compared to 53.3% for control participants. The *effort-weighted engagement* metric, which distinguishes immediate disengagement or success from effortful persistence, revealed that engagement trajectories diverged across groups as task complexity increased. Experimental participants maintained net positive weighted engagement across all 12 tasks while control participants crossed below the persistence threshold from Task 8 onward with growing disengagement. Frustration was also more prevalent in the control group: 76.3% of control participants exhibited at least one frustration bout compared to 55.7% of experimental participants, and task abandonment immediately following frustration was more than twice as common in the control group (7.9% of all tasks vs. 3.0%). Together these results provide strong support for rejecting H1 across both the frequency and quality of engagement.

Correctness. Experimental participants significantly outperformed control participants on both first-attempt correctness ($M = 0.635$ vs. $M = 0.539$) and first-intent correctness ($M = 0.781$ vs. $M = 0.582$). The control group outperformed the experimental group on Task 1 for both metrics, reflecting a genuine familiarity advantage in the early tasks rather than a reasoning advantage, as qualitative feedback confirmed. Experimental advantages emerged consistently from Task 4 onward on first-attempt correctness, and from Task 2 onward on first-intent correctness, with differences growing substantially across the final four tasks. H2 is therefore rejected, with the caveat that an initial interface adjustment cost accompanied the experimental interface’s longer-term reasoning benefits.

Confidence. Confidence showed the largest design effect of any outcome accounting for 15.2% of the variation in responses, with experimental participants reporting consistently higher confidence from Task 2 onward and a significant design–task interaction indicated that the gap widened with complexity. The largest differences emerged on Tasks 11 and 12, where experimental participants reported confidence nearly two points higher than control participants on the 7-point scale.

Anxiety. The main effect of design on anxiety was not statistically significant ($F(1, 110) = 2.51, \eta_G^2 = .014, p = .116$), but a significant design–task interaction indicated that anxiety trajectories diverged between groups as task complexity increased, with control participants reporting significantly higher anxiety on Tasks 2, 11, and 12. The non-significant main effect is best understood in the context of a floor effect identified upon inspection of affective ratings following task abandonment: 71 of 79 simultaneous 1/1 anxiety–confidence rating pairs came from control participants, and 88.6% of these followed a give-up. This artifact suppressed the control group mean toward the scale floor. However, as the outcome-conditional analysis in Table 4.17 shows, these low ratings are not completely straightforward; they appear to reflect a genuine affective state in which anxiety and self-efficacy collapsed together following repeated unproductive effort, a pattern discussed further in Section 4.6.4. The attenuation of the main effect is therefore itself meaningful: the control group produced the psychological conditions for that state more frequently. The significant interaction, which survived despite this attenuation, is consistent with a genuine anxiety divergence across tasks. H4 is therefore partially rejected, with the direction of effects contradicting the hypothesis but the main effect attenuated by a pattern of disengagement-driven affective collapse that was systematically more prevalent in the control group.

4.6.2 Interpretation

The primary mechanism under investigation was whether requiring explicit declaration of variable types, roles, and relationships (rather than assigning them implicitly through parameter slot population or test selection) would shift the majority of cognitive effort from procedural to conceptual. The results are consistent with this interpretation. In the experimental interface, every submission required participants to articulate what each variable *was* (its type and its analytical role) before the system would resolve what to do with it. This made semantic decisions inescapable: participants could not complete a task by selecting a plausible test and populating its argument slots, because there were no argument slots to populate. The qualitative record confirms that experimental participants were reasoning about variable meaning rather than navigating interface structure: comments describe active deliberation about whether a satisfaction variable was ordered or categorical, whether a lighting condition was repeated or input, and whether gender served as a control variable. Control participants, by contrast, described strategies oriented towards the interface itself (argument counting, description scanning, and systematic enumeration) rather than towards the data.

This distinction maps onto cognitive load theory’s distinction between extraneous and intrinsic cognitive load [132, 133]. Procedural demands that are unrelated to the target concept constitute extraneous load; their reduction frees working memory for the schema formation that produces transferable understanding. By automating test selection and preprocessing once the semantic design was declared, the experimental interface reduced the extraneous load of remembering test names, argument orders, and encoding conventions, while redirecting cognitive effort toward articulating analytical meaning. Importantly, this relocated cognitive effort rather than reduced it: the same number of decisions, made semantically upfront rather than

procedurally during parameter assignment. The learning trajectories reported in the Results are consistent with the behavioral signature of schema development: first-intent correctness and engagement improved across structurally diverse tasks in a pattern that cannot be explained by task-specific practice alone.

Both interfaces in this study were block-based environments that constrained syntactic variation through dropdown menus and pre-built blocks; neither required participants to type code. This was a deliberate methodological choice: the study’s goal was to examine how participants reasoned about and selected statistical analyses, not how they produced syntax. By holding the mode of interaction (block-based selection) constant across both groups, differences in first-intent correctness and engagement between groups can be more directly attributed to differences in statistical reasoning under the two conceptual frameworks being compared, rather than to differences in typing accuracy, syntax recall, or debugging. This design choice is not without limitation: because neither group typed code, it remains unknown whether the same pattern of results would persist in a free-typing environment, where confounding factors such as syntax errors, debugging demands, or familiarity with a programming language could interact with, or obscure, the conceptual differences observed here. Many participants in both groups noted initial unfamiliarity with the block-based paradigm, but this dissipated quickly: after the first few tasks, neither group’s feedback suggested the interface itself was a source of friction, with participants describing it as functional, intuitive and workable even when the statistical content remained challenging. This relatively rapid familiarization was likely supported by the intentional constraints of the design and allowed participants to focus on the task content.

A consistent theme across the correctness and engagement data is that the experimental interface imposed an initial adjustment cost that gave way to sustained performance advantages, while the control interface supported early performance through familiarity but provided diminishing returns as complexity increased. This crossover pattern is corroborated by role-specific learning trajectories: experimental participants improved across the repeated-measures tasks and the covariate tasks, while control participants showed no significant trajectory on either role type, perhaps due to changing syntax requirements as the complexity increased. The pattern suggests that by requiring explicit declaration across structurally diverse tasks, the experimental interface appears to have supported the gradual internalization of a conceptual framework, one in which variables have types and roles that determine analytical choices, that participants applied with increasing fluency to new task configurations. Experimental participants’ end-of-study comments are consistent with this reading, with several describing a moment of transition from confusion to fluency, using the language of things “intuitive”(P6, P13, P20, P87 and P102), “making a lot of sense”(P104), or “seeming natural”(P30), and characterizing tasks as being “like a puzzle”(P17) and “easier to understand over time”(P50) relying on “problem-solving”(P95) rather than interface navigation.

The substantially larger design effect for first-intent correctness ($\eta_G^2 = .113$) compared to first-attempt correctness ($\eta_G^2 = .022$) speaks to the nature of the experimental interface’s reasoning demands. The block-based experimental interface encouraged an incremental building style: participants added variables one at a time, received per-block feedback, and refined their design iteratively. This inflated attempt counts

even for correct underlying reasoning, because a first submission of an incomplete workspace registered as incorrect even if all submitted elements were individually accurate. First-intent correctness, by isolating each reasoning decision the first time it was made, removes this artifact and captures underlying reasoning quality more directly. For the research question this study was designed to address, first-intent is therefore the more promising primary metric.

4.6.3 Generalizability

Participants were recruited exclusively from STEM degree programs at a single university, predominantly Computer Science, Computer Engineering, and Cybersecurity. This homogeneity limits generalizability to populations with different academic backgrounds, particularly those in social, behavioral, or biological sciences where statistical reasoning is similarly central but computational familiarity typically differs. Research on statistical anxiety has found stronger gender effects and higher baseline anxiety in non-STEM and female-majority samples [134, 103], and the present sample’s male-dominated composition (79 male, 31 female among completers) and STEM orientation may have dampened observable gender effects in both directions. The general absence of gender differences in this study should therefore not be interpreted as evidence that gender plays no role in interactions with statistical software more broadly. The exploratory disability findings, which tentatively suggest the experimental interface may have been particularly supportive for participants with disabilities, are based on a subsample of 20 participants (8 experimental, 12 control) and should be treated as hypothesis-generating. Replication with a larger and adequately powered sample across a broader range of disability types is needed before drawing firm conclusions.

The block-based interface was a deliberate design choice that attempted isolated statistical reasoning from programming ability. The tradeoff is that findings may not transfer directly to free-form text coding environments. In practice, statistical software is used in contexts that require syntax recall, error debugging, and iterative code execution, demands that the present design intentionally removed. Whether the reasoning advantages observed here would persist, attenuate, or amplify when the same semantic framework is implemented in a free-form coding environment where syntactic demands reintroduce extraneous load and semantic feedback may not be available is an open question. The present results speak to what participants understand about their data and analysis; they do not speak to whether that understanding would survive the additional demands of a live coding context.

The 12-task sequence covered a broad range of statistical and machine learning analyses, from simple correlation through mixed ANCOVA and logistic regression, providing reasonable construct coverage of applied statistics. However, tasks were necessarily simplified relative to real research analyses: datasets were small and clean, variable roles and task descriptions were generally unambiguous by design, and preprocessing requirements were constrained. Real-world analyses involve messy data, ambiguous variable structures, and domain knowledge that the present design could not incorporate. The observed correctness advantages may therefore underestimate the benefits of explicit semantic declaration in more complex, ecologically realistic

settings where the cognitive cost of remembering argument orders and encoding conventions is higher.

The study was conducted in a single session with no follow-up measurement. The learning trajectories observed across the 12 tasks are suggestive of conceptual development, but without a retention or transfer assessment it is not possible to determine whether the understanding experimental participants appeared to develop persisted beyond the session or transferred to tasks outside the studied sequence. The fixed task ordering also means that learning effects and task complexity effects are confounded: it is not possible to determine how much of the trajectory finding reflects genuine schema development versus accumulated familiarity with the specific interface mechanics across the session. A counterbalanced or longitudinal design would be needed to disentangle these contributions.

Anxiety and confidence were assessed using single-item 7-point Likert scales. As discussed in Section 4.3.10, established multi-item instruments measure trait-level statistical anxiety and self-efficacy as stable dispositions and were not designed for repeated task-specific measurement across 12 tasks; single-item scales were therefore adopted to capture moment-to-moment affective responses without inducing response fatigue. Several design decisions were implemented to reduce satisficing and mis-clicks: scales were presented one at a time with the confidence scale advancing automatically after a 300ms delay following anxiety selection, response options used full descriptive labels unique to each construct (e.g., *No anxiety at all* vs. *No confidence at all*) to discourage mechanical click-through, radio buttons and their labels were large and the area was fully clickable, reducing the motor precision required to register a selection [46]. Participants could review and change both ratings before advancing. However, one experimental participant left a task-7 comment: “Accidentally clicked large amount of anxiety. I actually had only a small amount this time”, perhaps unaware that they could change their selection before moving on. This suggests some measurement error remains. Ultimately, these measures demonstrated convergent validity and matched qualitative patterns, but lack the psychometric depth of validated multi-item instruments [139, 29, 56, 44] and may not fully capture the complexity of these constructs at the trait level.

4.6.4 Implications

Implications for statistical software design. The present findings support a concrete design principle: making semantic decisions (i.e. variable type, analytical role, and structural relationships) explicit and required, rather than implicit and inferred, improves both the quality of statistical reasoning and the affective experience of the analytical process. This principle is not specific to the experimental implementation evaluated here. R’s formula interface moved in this direction by asking users to specify variable relationships rather than invoke algorithms directly, but leaves type distinctions, repeated structure, and encoding behavior implicit and reliant on symbolic syntax. Wrapper packages such as *ez* in R and *pingouin* in Python reduce surface inconsistency but do not eliminate the cognitive burden of test selection. The present results suggest that the remaining gap, the semantics that cannot be inferred from data alone, is precisely where the largest performance and affective benefits lie.

Implications for statistics education. The learning trajectory findings carry pedagogical implications that extend beyond software design. The pattern of improving role and type accuracy across structurally diverse tasks suggests that explicit semantic declaration may function as a form of deliberate practice for the conceptual distinctions that applied statistics requires. Students who repeatedly commit to a variable type or analytical role, and receive specific feedback when that commitment is incorrect, appear to develop a more robust conceptual framework (measured here using first-intent correctness) than students who assign variables to parameter slots, where the slot structure itself conveys the role without requiring the student to articulate it. This suggests a potential pedagogical application: requiring students to complete a structured semantic specification of their analysis prior to implementation, in any software environment, may support deeper conceptual development than implementation-first approaches that allow procedural performance to substitute for conceptual understanding.

The exploratory disability finding, though preliminary, warrants attention from an educational and accessibility perspective. The tentative pattern of lower anxiety, higher confidence, and trending reductions in give-up rate for experimental participants with disabilities is consistent with the hypothesis that explicit structure may reduce the cognitive overhead associated with navigating ambiguous interface demands. This represents a promising direction for accessibility-focused statistical software research and should be replicated with an adequately powered sample before drawing firm conclusions.

Implications for the study of statistical anxiety. The affective data in this study suggest a reframing of what statistical software design should aim to achieve with respect to anxiety. The goal is not the elimination of anxiety but the preservation of its productive function: signaling difficulty without collapsing the learner’s sense of agency. As Table 4.17 shows, the experimental interface produced an anxiety arc that makes functional sense (highest when participants gave up without engaging at all, moderate during exploration, and lowest after success) while confidence moved in the opposite direction, growing as participants progressed from confusion toward mastery. This is optimal anxiety being present when something genuinely exceeds current understanding and resolving as understanding develops, supporting Macher and colleagues’ account of statistics anxiety as, in their words, a “blessing in disguise” [85] with a positive influence on performance.

The control group broke this arc. Anxiety after complete disengagement was virtually indistinguishable from anxiety after success, while confidence remained near the floor across all give-up states regardless of effort invested: the same low anxiety value described two opposite psychological states depending on which group produced it. We term this pattern the *anxiety-to-apathy transition*. Notably, the data suggests that once confidence collapses toward the floor, effort may no longer feel connected to outcome, and anxiety, which depends on the outcome still mattering, has nothing left to attach to. Apathy, on this view, emerges not because anxiety became unbearable, but because there was nothing left to be anxious *about*, distinct from anxiety reduction through success, and from the moderate, still-resolvable anxiety of confusion. Several control participants described this directly: “after a while apathy took over and I didn’t have the energy

to keep trying”(P35); “over time, learned helplessness instills and it’s hard to overcome it”(P29); “very enthusiastic at first but start to give up as nothing more gets progress”(P98); “no idea so no anxiety or confidence”(P16); “the tasks keep getting more complex. I don’t even attempt at this point so my anxiety is not high”(P69); “my anxiety did not exceed moderate levels as I have the option to give up and move on”(P45). Several of these accounts name the absence of confidence, not overwhelming anxiety, as the reason anxiety subsided, consistent with a confidence-driven account. Similar accounts emerged in Study 1 (Chapter 3). This aligns with Seligman’s [124] account of learned helplessness, in which repeated uncontrollable failure produces withdrawal from responding rather than ongoing distress, once the organism stops perceiving a connection between its own effort and the outcome.

The implication for statistics anxiety research and software design is twofold. For measurement, low anxiety ratings in performance contexts are not straightforwardly interpretable without knowing the participant’s engagement state at the time of rating: the same value can reflect genuine ease or genuine apathy, and the layout of the interface may inadvertently capture certain affective states (as the give-up button location documented in the Results illustrates) making the confidence score necessary to distinguish between them. Conditioning analyses on task outcome, as in the present work, is one approach; designing rating procedures sensitive to the reason for low anxiety is another. For design, what matters is not whether an interface reduces anxiety on average, but whether it preserves the productive relationship between anxiety and engagement, keeping anxiety informative rather than overwhelming, and ensuring that effort can resolve it. The experimental interface appears to have achieved this. The control interface did not, not because it produced too much anxiety, but because it produced anxiety that effort could not address. Ultimately, the findings suggest that the sea of statistics anxiety is not the obstacle, it is the environment. What determines whether learners sink or swim is whether the tools they are given make effort meaningful.

Chapter 5

Declarative Layer for Statistical and Machine Learning Workflows

5.1 Introduction

The preceding studies examined how people interact with statistical and machine learning interfaces, and what kinds of design decisions affect their ability to correctly specify and interpret analyses. A consistent theme across both is that users struggle not with computation itself, but with the gap between their understanding of their data and the interface’s expectations. Most statistical libraries are organized around analyses: a user must already know what test they want in order to ask for it. This places the burden of methodological knowledge at the point of entry, before the system has any information about what the user is actually trying to understand.

This chapter describes a design response to that finding. Rather than organizing a library around analyses as the primary object, we propose organizing it around the research design itself, specifically, around a declarative description of the variables in a dataset, their types, their roles in the analysis, and the relationships among them. We call this object a *ModelDesign*. Once a *ModelDesign* is specified, it can serve as a shared substrate for routing to appropriate statistical tests, fitting machine learning models, and, in the future, generating trustworthy visualizations and interpretable output summaries.

The result is a two-layer architecture. The lower layer is an existing statistics library (discussed in Chapter 6), built around the naming conventions and analytical categories that emerged from our first study. That library remains largely unchanged with some additional analyses implemented. The upper layer is a declarative framework, the *ModelDesign* system, that wraps and routes to the lower layer. The system is designed to be extensible: additional analyses, model families, variable types, and encoders can be added over time without restructuring the core.

One explicit design boundary is worth stating at the outset. This system assumes that data has already

been brought into Tidy format; that is, each row represents an observation, each column a variable, and each cell a single value [145]. This is not an arbitrary restriction. The same precondition underlies the dominant conventions of scripting-based statistical computing: R’s formula interface maps variable names directly to columns using Wilkinson notation [148], the Tidyverse is built around this structure by design, and Python-based libraries such as *statsmodels* and *scikit-learn* similarly expect a columns-as-variables representation. Data wrangling is its own complex domain, and the usability of tools for that process warrants separate empirical investigation. By treating Tidy format as an entry condition, we make the scope of this system tractable and concentrate its contribution on the modeling layer, where the evidence from our studies most directly applies.

5.2 Declare a Distinct Identifier

We use the term identifier to broadly refer to the unique individual or entity being studied, encompassing a range of terminology across statistical fields. The library’s earlier test-based implementation, evaluated in Study 1, accommodated both wide and long data formats, allowing identifiers to sometimes be inferred or omitted depending on how the data was structured. The declarative framework presented in Study 2 of this dissertation instead expects data in tidy format, in which each row represents a single observation and each column represents a single variable. Within this format, an explicit subject identifier becomes essential whenever repeated measurements are collected from the same subjects, since multiple rows will correspond to a single subject and the identifier is the only mechanism by which those rows can be correctly mapped back to one another. Without it, the system has no way to determine which observations belong to the same subject, making accurate grouping, aggregation, or within-subject comparison impossible. For example, subject identifiers enable the calculation of summary statistics, such as averaging repeated measures for each individual.

Within the statistical library, the subject identifier is a foundational component for organizing and managing data, and is required in any repeated measures contexts to align, group, reshape, and merge data correctly for statistical tests, charts, and reports. Behind the scenes in General Linear Model contexts [96], the identifier column is encoded [24] using a *OneHotEncoder* [107] with N_s levels, where N_s represents the total number of unique subjects. It provides a consistent reference point for subject-level operations and ensures reliable handling of data across repeated measures, within-subject designs, and other analyses requiring subject-level coherence.

The current framework supports a single, flat identifier per subject. Future extensions to Declarative Data Science may introduce *nested* identifiers, capable of representing hierarchical or multilevel data structures, such as students nested within classrooms, or patients nested within clinics, enabling the declarative framework to extend naturally to multilevel and hierarchical modeling techniques. In anticipation of this possibility, the identifier variable constructor would be better named generically as *CreateIdentifierVariable* rather than *CreateSubjectIdentifierVariable* (as seen in Study 2), allowing the same role to be extended to

nested or multilevel identifiers in future versions of the library without requiring a change to the underlying naming convention.

5.3 Declare Input Variable(s)

We use the term “input” to encompass a range of concepts across various statistical fields, all generally referring to similar ideas. Common terms include independent variables, predictors, features, explanatory variables, regressors, factors, covariates, and exposure variables. Inputs in this system can vary in role: they may be Input, Repeated, or Control variables, and they can take on different data types: Categorical, Numerical, or Ordinal. This terminology allows us to capture the diversity of variables and data structures found in statistical analysis. Some analyses and models do not require any outputs, so input-only designs are accepted by *GetTest* and *GetModel*.

Numerical-(Input or Control) This option allows you to add a numerical independent variable or covariate to your model before performing inferential analysis, building a model, or creating a chart. In this context, the numerical independent variable is treated as a continuous predictor, and its relationship with the dependent variable is modeled directly to assess how changes in the predictor influence the outcome. In *GetTest*, this variable would be used with its raw values and would not be standardized by default for GLM analyses. However, in *GetModel*, it would be normalized using *StandardNormalizer* by default.

Categorical-(Input, Repeated, or Control) This option allows you to add a categorical independent variable (i.e. a between-subjects factor, a within-subjects factor, or a categorical covariate) to your model before performing inferential analysis or creating a chart. The encoding behavior described here applies identically regardless of which of these three roles is declared; role does not alter how the variable is encoded, only which statistical procedure is selected once encoding is complete. In *GetTest*, this variable would be encoded using a *ReferenceEncoder* by default for GLM analyses, known as dummy encoding, where the reference and level order can be set. However, when used in *GetModel*, it would be encoded using a *OneHotEncoder* by default.

Ordinal-(Input, Repeated, or Control) This option allows you to add an ordinal independent variable to your model. The encoding behavior described here applies identically regardless of which of these three roles is declared. In *GetTest* and *GetModel*, this variable would be encoded using an *IntegerEncoder*. However, if the data is text data, a level order must also be defined using *SetLevelOrder*.

5.4 Declare Input Interaction(s)

Users may also declare an interaction between two or more inputs directly, independent of whether the corresponding main effects have been declared on their own. As in Wilkinson notation, where *A:B* specifies

the interaction term alone and $A*B$ expands to the interaction plus its lower main effects ($A + B + A:B$) [148], the declarative framework distinguishes between these two cases through a single boolean argument rather than two separate operators. By default, declaring an interaction between two inputs automatically includes their lower-order terms, mirroring the behavior of $*$ regardless of whether those terms were separately declared elsewhere. A user may instead request the interaction term in isolation to mirror the behavior of $:$ by explicitly setting an argument such as *include_lower_order=false* (naming to be determined) when constructing the interaction. As with the other declarative constructs in this chapter, the role and type of each individual input (Categorical, Numerical, or Ordinal) are already known from their individual declarations, so the interaction term’s encoding is derived automatically from those declarations rather than requiring separate specification.

5.5 Declare Output Variable(s)

We use the term “output” to encompass a range of concepts across various statistical fields, all generally referring to similar ideas. Common terms include dependent variables, outcome variables, response variables, regressands, measured variables, target variables, and criterion variables. Outputs can vary in type: they may be Categorical, Numerical, or Ordinal, depending on the nature of the data and the analysis. Unlike inputs, output variables cannot be declared with the Repeated or Control roles, since these roles only apply to variables used to structure or adjust for relationships among predictors, not to the variable being predicted or tested. Some analyses do not require any inputs, so output-only designs are accepted by *GetTest*, but *GetModel* requires inputs to be present.

Numerical Output This option allows you to add a numerical dependent variable to your model. Numerical outputs are never standardized or otherwise transformed by default, in either *GetTest* or *GetModel*, since the scale of the outcome variable is generally meaningful to its interpretation and altering it would change how results are reported and understood.

Ordinal Output This option allows you to add an ordinal dependent variable to your model. If the underlying data is text (e.g., “Low,” “Medium,” “High”), a level order must be defined using *SetLevelOrder*, after which the variable is encoded using an *IntegerEncoder*. If the underlying data is already numeric or integer-valued and no explicit encoding is requested, the variable is passed through unmodified in *GetTest*, since rank-based tests perform their own internal ranking and do not require pre-encoded values. In *GetModel*, the variable is encoded using an *IntegerEncoder* by default, since predictive models require a numeric representation of the outcome regardless of whether the original values were already numeric.

Categorical Output This option allows you to add a categorical dependent variable to your model. In *GetTest*, this variable is passed through unmodified, since the relevant tests construct their own count or

contingency tables directly from the variable’s category labels. In *GetModel*, this variable is encoded using a *OneHotEncoder* by default; alternatively, a *LabelEncoder* may be used, which applies integer encoding but is reserved specifically for output variables, distinguishing it from the *IntegerEncoder* used for ordinal data, since label encoding assigns arbitrary integer codes to nominal categories without implying any order among them.

5.6 Declare an Analytical Goal

Once a variable’s type, role, and any relationships have been declared, the user selects an entry point corresponding to their analytical goal: *GetSummary* for descriptive statistics, *GetTest* for inferential statistics, or *GetModel* for predictive modeling. A potential future entry point, *GetChart*, could extend this same principle to data visualization. In this system, the user first describes what their data is, and then asks what they wish to do with it. As such, selecting one of these entry points is not equivalent to selecting a statistical procedure; it is a declaration of analytical intent, whether the user wishes to describe their data, test a relationship within it, or predict future observations from it. The specific procedure used to accomplish that goal, whether a particular test, model, or visualization, is decided deterministically from the semantic structure already declared, as described throughout the next chapter. The analytical goal and the semantic structure of the data are both decisions intrinsic to understanding one’s research question, while the specific procedure is not.

Chapter 6

Implementation and Mathematical Foundations

This chapter documents the mathematical procedures underlying *GetSummary*, *GetTest*, and *GetModel*: the three entry points through which a user, having declared their variables by type and role, receives an appropriate descriptive statistic, inferential test, or predictive model automatically, without needing to name the procedure themselves. The chapter is organized around this same structure. Each major section documents the individual procedures available under one of these three entry points, followed by the routing logic that determines, from a user's declarations alone, which of those procedures is selected. The routing logic is, in a real sense, the point of this library: the procedures themselves are drawn from established statistical practice, but the automatic mapping from a user's data to the correct procedure is the library's central contribution, and the reason the underlying mathematics needed to be documented as thoroughly as it is here.

That thoroughness deserves its own explanation, since a reader may reasonably ask why a dissertation would re-derive equations that are, in many cases, already well established elsewhere. Three reasons motivate it. First, and most practically, this documentation was a necessary step in building the library. Every routing decision described in this chapter corresponds to a real implementation, and implementing a procedure requires first knowing, precisely, what it computes. Second, considerable effort went into formulating these equations in a single, consistent standard, generalized as fully as possible rather than limited to the simplest case of a given design, and accompanied throughout by their corresponding post hoc procedures, effect sizes, corrections, and assumption tests. Third, and most importantly, a substantial portion of the equations in this chapter (and many of those not included in this writeup) were considerably harder to track down, verify, or correctly generalize than their routine use in published research would suggest. Some were scattered across decades-old literature that presented only a basic case (e.g., a one-way simple design) while leaving the

generalization to n -way, mixed, or unbalanced designs implicit or unaddressed. Some sources presented an equation without enough surrounding explanation to confirm it was being interpreted correctly. For example, in at least one case, a published equation contained an error that had, as far as could be determined, only ever been corrected in an informal exchange between a practitioner and the original author [32]. Centralizing these equations in one place, verified against both the statistical literature and this library's actual source code, is intended as a resource in its own right, independent of its role in justifying the library's design.

This chapter aims to provide a detailed understanding of the tools and methodologies available within the library, showcasing its potential versatility and utility in addressing a broad spectrum of statistical and data science challenges. While considerable effort has been made throughout this chapter to verify each formula against its original source or a reliable secondary reference, the reader is encouraged to independently verify the equations presented here, particularly before relying on them for implementation or replication. Several of the procedures described draw on formulas and corrections distributed across decades of literature and across several differing notational conventions, and it is possible that errors remain despite this effort.

6.1 Descriptive Statistics

GetSummary (naming to be determined) is the system's entry point for descriptive statistics. Once a user has declared the variables of interest, by specifying their type (Numerical, Ordered, or Categorical) and role (Input, Output, Repeated, Control, or Identifier), *GetSummary* returns the descriptive measures appropriate to those declarations automatically. A numerical variable, for example, returns measures of central tendency, variability, and shape, while a categorical variable returns frequency counts and proportions instead.

Role declarations shape not just which variable is summarized, but how. If a categorical variable is declared with the Input role, it acts as a grouping variable, and any associated Output variable is summarized separately within each group rather than across the full sample. The specific measures reported in this case are limited to those most relevant to the specified design rather than a complete breakdown.

```
DataFrame frame
frame:Load("data.csv")
CategoricalVariable group = frame:CreateCategoricalInputVariable("group")
NumericalVariable score = frame:CreateNumericalOutputVariable("score")
ModelDesign design
design:AddVariable(group)
design:AddVariable(score)
StatisticalSummary summary = frame:GetSummary(design)
output summary:ToText()
```

Figure 6.1: Experimental interface pseudocode for generating a statistical summary.

When two or more output variables are declared together without any input, *GetSummary* instead returns a measure of association between them, selected according to the variable types involved, for example, Pearson correlation for two numerical variables, or Cramér's V for two categorical variables. In either case, the user is not required to know in advance which descriptive statistic or formula applies to their data; the appropriate output follows automatically from the type and role declarations already made.

6.1.1 Measures of Central Tendency

These measures summarize a dataset by identifying a central or typical value around which other data points are distributed. They help describe where most values in a distribution are located.

Mean

The average value of numerical data.

Population Mean (μ)

x_i represents each value in the population

N is the total number of values in the population

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (6.1)$$

The population mean (often denoted as μ or μ_x) is the average of all values in an entire population. It represents the central value of the population data.

Sample Mean ($\bar{x}$)

x_i represents each value in the sample

n is the total number of values in the sample

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (6.2)$$

The sample mean (often denoted as $\bar{x}$) is the average of all values in a sample taken from the population. It often serves as an estimate of the population mean. It is only an approximation and will vary depending on the sample chosen.

Median

The middle value when the data is ordinal or numerical.

Population Median (μ_m)

x_i represents each value in the ordered population dataset

N is the total number of values in the population

If N is odd:

$$\mu_m = x_{\left(\frac{N+1}{2}\right)} \quad (6.3)$$

If N is even:

$$\mu_m = \frac{x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}}{2} \quad (6.4)$$

The population median (often denoted as μ_m , M , Med or Mdn) is the value that divides the entire population data set into two equal halves. It represents the central value when all values in the population are ordered from smallest to largest.

Sample Median ($\tilde{x}$)

x_i represents each value in the ordered sample dataset

n is the total number of values in the sample

If n is odd:

$$\tilde{x} = x_{\left(\frac{n+1}{2}\right)} \quad (6.5)$$

If n is even:

$$\tilde{x} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \quad (6.6)$$

The sample median (often denoted as $\tilde{x}$, M , Med or Mdn) is the median of a sample taken from the population. Like the population median, it divides the sample data into two equal parts when ordered, providing an estimate of the central tendency of the population.

Mode

The most frequently occurring value in ordinal or categorical data.

Population and Sample Mode ($mode$)

$f(x)$ is the frequency of x in the dataset.

$$mode = \operatorname{argmax}_x f(x) \quad (6.7)$$

The mode is determined by identifying the value x that appears most frequently in the dataset, it is calculated the same way for both population and sample datasets.

6.1.2 Measures of Variability

These measures indicate how much the data values vary or spread out from the center. They provide insights into the range and consistency of the data.

Variance

The average of the squared differences from the mean, showing how spread out the data is.

Population Variance (σ^2)

x_i represents each value in the population

μ is the population mean from equation (6.1)

N is the total number of values in the population

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (6.8)$$

The population variance (often denoted as σ^2 or σ_x^2) calculates the variance of a population and uses N in the denominator.

Sample Variance (s^2)

x_i represents each value in the sample

$\bar{x}$ is the sample mean from equation (6.2)

n is the total number of values in the sample

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (6.9)$$

The sample variance (often denoted as s^2) estimates the variance of a sample and uses $n-1$ in the denominator to account for bias in estimating the population variance from a sample.

Standard Deviation

The square root of variance, indicating the average distance of each data point from the mean.

Population Standard Deviation (σ)

σ^2 is the population variance from equation (6.8)

$$\sigma = \sqrt{\sigma^2} \quad (6.10)$$

The population variance (often denoted as σ or σ_x) calculates the the spread of data points around the mean of a population.

Sample Standard Deviation (s)

s^2 is the sample variance from equation (6.9)

$$s = \sqrt{s^2} \quad (6.11)$$

The sample standard deviation (often denoted as s) estimates the spread of data points around the mean of a sample and accounts for bias in estimating the population standard deviation from a sample.

Range

The difference between the maximum and minimum values.

Interquartile Range (IQR)

The range of the middle 50% of the data, or the difference between the first and third quartiles.

6.1.3 Measures of Shape

These measures describe the overall shape and symmetry of the data distribution. They indicate how the data values are distributed relative to the mean and can signal any skewness or unusual concentration of values.

Skew

Describes the asymmetry of the data distribution. Positive skew means a tail on the right, while negative skew means a tail on the left. [64]

Population Skewness (γ_1)

x_i represents each value in the population

μ is the population mean from equation (6.1)

σ is the population standard deviation from equation (6.10)

N is the total number of values in the population

$$\gamma_1 = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^3}{\sigma^3} \quad (6.12)$$

The population skewness (often denoted as γ_1 or $skew(x)$) measures the asymmetry of data around its mean.

Sample Skewness (g_1)

x_i represents each value in the sample

$\bar{x}$ is the sample mean from equation (6.2)

s is the sample standard deviation from equation (6.11)

n is the total number of values in the sample

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3 \quad (6.13)$$

The sample skewness (often denoted as g_1 or $skew(x)$) measures the skewness of a sample and corrects for bias in small samples by adjusting for the sample size.

Kurtosis

Measures the “tailedness” of the distribution; high kurtosis indicates heavy tails, and low kurtosis indicates light tails. [64]

Population Kurtosis (γ_2)

x_i represents each value in the population

μ is the population mean from equation (6.1)

σ is the population standard deviation from equation (6.10)

N is the total number of values in the population

$$\gamma_2 = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^4}{\sigma^4} \quad (6.14)$$

The population kurtosis (often denoted as γ_2 or κ) measures how much of the data falls within the tails.

Sample Kurtosis (g_2)

x_i represents each value in the sample

$\bar{x}$ is the sample mean from equation (6.2)

s is the sample standard deviation from equation (6.11)

n is the total number of values in the sample

$$g_2 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{s^4} - 3 \quad (6.15)$$

The sample kurtosis (often denoted as g_2 or $\hat{\kappa}$) corrects for bias by adjusting for the sample size. The subtraction of 3 in the sample kurtosis formula centers the value to compare kurtosis against a normal distribution, where excess kurtosis of 0 indicates "mesokurtic" behavior, like a normal distribution.

6.1.4 Measures of Association

These types of measures assess the strength and direction of the relationship between two variables and includes statistics that indicate how the two variables move together.

Covariance

Covariance indicates the direction of a linear relationship between two variables (positive or negative), but not its strength or consistency.

Population Covariance (σ_{XY})

x_i represents each value in X

y_i represents each value in Y

μ_X is the population mean of X

μ_Y is the population mean of Y

N is the number of paired observations in the population

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y) \quad (6.16)$$

The population covariance (often denoted as σ_{XY} or $Cov(X, Y)$) measures the direction of the linear relationship between two variables, showing how changes in one tend to be associated with changes in the other.

Sample Covariance (s_{xy})

x_i represents each value in x

y_i represents each value in y

$\bar{x}$ is the sample mean of x

$\bar{y}$ is the sample mean of y

n is the number of paired observations in the sample

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (6.17)$$

The sample covariance (often denoted as s_{xy} or $Cov(x, y)$) uses $n - 1$ in the denominator to produce an unbiased estimate of the population covariance from sample data.

Correlation for Numerical Data

Correlation is a standardized form of covariance that quantifies both the direction and strength of the linear relationship between two variables, ranging from -1 (perfect negative) to 1 (perfect positive).

Pearson's Correlation Coefficient (r) [106]

x_i represents each value in x

y_i represents each value in y

$\bar{x}$ is the sample mean of x

$\bar{y}$ is the sample mean of y

n is the sample size; $n_x = n_y$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right) \left(\sum_{i=1}^n (y_i - \bar{y})^2\right)}} \quad (6.18)$$

The formula is identical for population and sample, differing only in notation: the population correlation, ρ or $\rho(X, Y)$, uses population means μ_X , μ_Y , and count N , while the sample correlation, r or $r(x, y)$, uses sample means $\bar{x}$, $\bar{y}$, and count n .

Correlation for Ordinal Data

Spearman's Correlation Coefficient (r_s) [127]

x is the input variable

y is the output variable

n is the sample size: $n_x = n_y$

d_i is the difference in ranked values: $R(x_i) - R(y_i)$

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (6.19)$$

This coefficient measures the strength and direction of a rank-based correlation. r_s ranges from -1 to 1. A perfect negative correlation would be indicated by -1, and a perfect positive correlation would be indicated by 1. If r_s is 0, this would indicate there is no correlation at all. *Note: Not yet implemented.*

Tie-Corrected Spearman Rank Correlation Coefficient (r_s) [128]

t_i is the i^{th} tie length in x

t_j is the j^{th} tie length in y

$$T_x = \frac{n(n^2 - 1) - \sum (t_i^3 - t_i)}{12} \quad (6.20)$$

$$T_y = \frac{n(n^2 - 1) - \sum (t_j^3 - t_j)}{12} \quad (6.21)$$

$$r_s = \frac{T_x + T_y - (\sum_{i=1}^n d_i^2)}{2\sqrt{T_x * T_y}} \quad (6.22)$$

This coefficient measures the strength and direction of a rank-based correlation while correcting for ties between ranks. r_s ranges from -1 to 1 it and can be interpreted in the same way as the uncorrected r_s from equation (6.19). *Note: Not yet implemented.*

Spearman Rank Correlation Coefficient Via Pearson's Formula (r_s)

This method would produce the same coefficient as equation (6.19) but using an alternative technique. It applies Pearson's correlation formula (6.18) to the ranked values of the samples. This method is often used in statistical software in order to utilize much of the same code for the two tests. This equation generally gives a sufficient approximation of Spearman's Correlation when ties are present as the ties are averaged during the ranking process. However, if there are a substantial number of ties, using the tie-corrected equation (6.22) would provide a more accurate result.

$R(x)$ is the input variable x ranked

$R(y)$ is the output variable y ranked

$\bar{R}(x)$ is the sample mean of $R(x)$

$\bar{R}(y)$ is the sample mean of $R(y)$

n is the sample size; $n_x = n_y$

$$r_s = \frac{\sum_{i=1}^n (R(x)_i - \bar{R}(x)) (R(y)_i - \bar{R}(y))}{\sqrt{\sum_{i=1}^n (R(x)_i - \bar{R}(x))^2} \sqrt{\sum_{i=1}^n (R(y)_i - \bar{R}(y))^2}} \quad (6.23)$$

This coefficient measures the strength and direction of a rank-based correlation. r_s ranges from -1 to 1. A perfect negative correlation would be indicated by -1, and a perfect positive correlation would be indicated by 1. If r_s is 0, this would indicate there is no correlation at all.

Correlation for Categorical Data

Cramér's Correlation Coefficient (ϕ) [28]

n is the sample size: $n_x = n_y$

χ^2 is the statistic calculated using equation (6.53)

$$\phi = \sqrt{\frac{\chi^2}{n}} \quad (6.24)$$

This coefficient measures association between two binary variables, used for 2x2 contingency tables. It ranges from -1 to 1, with values further from 0 indicating a stronger association and 0 indicates no association. A value of |0.1| is considered a small effect, |0.3| a medium effect, and |0.5| a large effect.

Cramér's Correlation Coefficient (V) [28]

n is the sample size: $n_x = n_y$

k is the number of categories in the first sample

m is the number of categories in the second sample

χ^2 is the statistic calculated using equation (6.53) or (6.55)

$$\phi^2 = \frac{\chi^2}{n} \quad (6.25)$$

$$V = \sqrt{\frac{\phi^2}{\min(k-1, m-1)}} \quad (6.26)$$

Unlike the phi coefficient (ϕ), V ranges from 0 to 1 and reflects only the strength of association. This loss of directionality is a consequence of the generalization: sign has no meaningful interpretation in tables larger than 2×2 .

Bias-Corrected Cramér's Correlation Coefficient ($\tilde{V}$) [9]

This corrected coefficient adjusts for the fact that Cramér's V can be biased when the sample size is small or when the contingency table is sparse.

$$\tilde{\phi}^2 = \max(0, \phi^2 - \frac{(k-1)(m-1)}{n-1}) \quad (6.27)$$

$$\tilde{k} = k - \frac{(k-1)^2}{n-1} \quad (6.28)$$

$$\tilde{m} = m - \frac{(m-1)^2}{n-1} \quad (6.29)$$

$$\tilde{V} = \sqrt{\frac{\tilde{\phi}^2}{\min(\tilde{k}-1, \tilde{m}-1)}} \quad (6.30)$$

It ranges from 0 to 1, with higher values indicating a stronger association. If Cramér's V is calculated using the χ^2 from equation (6.55), Yates' correction has already been applied, adding another bias correction would not significantly change the outcome. *Note: Not yet implemented.*

6.1.5 Future Additions

The measures described in this section represent the descriptive statistics most commonly used and currently implemented within the library, but are not an exhaustive account of all descriptive measures available in the broader statistical literature. Additional measures such as the geometric and harmonic means, the trimmed mean, the mean absolute deviation, the coefficient of variation, Kendall's correlation coefficient (r_τ), point-biserial correlation coefficient (r_p), and the correlation ratio (r_η) for mixed categorical-numerical associations

are well-established but are not currently implemented within the library, and represent possible extensions to its descriptive statistics functionality.

6.2 Inferential Statistics

GetTest (naming to be determined) is the system's entry point for inferential statistics. As with descriptive statistics, the user declares variable type (Numerical, Ordered, or Categorical) and role (Input, Output, Repeated, Control, or Identifier), and *GetTest* selects and executes the appropriate test automatically. A categorical Input paired with a numerical Output, for example, routes to an independent samples *t*-test or one-way ANOVA depending on the number of groups, while two numerical variables declared without an Input route to a correlation test.

Role declarations also determine which variant of a test is used. A Repeated role shifts the design from between-subjects to within-subjects, for example from an independent samples *t*-test to a paired *t*-test, or from a between-subjects to a repeated-measures ANOVA. A Control variable similarly shifts the test to one that can account for it, for example from a one-way ANOVA to an ANCOVA.

```
DataFrame frame
frame:Load("data.csv")
CategoricalVariable group = frame:CreateCategoricalInputVariable("group")
NumericalVariable score = frame:CreateNumericalOutputVariable("score")
ModelDesign design
design:AddVariable(group)
design:AddVariable(score)
StatisticalTest test = frame:GetTest(design)
StatisticalTestResult result = test:Run()
output result:ToText()
```

Figure 6.2: Experimental interface pseudocode for generating an inferential test.

Before execution, *GetTest* checks each test's assumptions (e.g., normality, homogeneity of variance, sphericity) against the declared data. Where an assumption is violated, an appropriate correction is applied where one exists, for example Welch's *t*-test in place of Student's *t*-test under unequal variances, or Kruskal-Wallis in place of one-way ANOVA under non-normality. Where no correction is available, the violation and a suggested alternative are reported alongside the test result.

Throughout this section, significance tests are reported as a probability $P(\cdot)$ relative to the relevant sampling distribution (e.g., t , F , χ^2 , Z , or q), following the convention of the specific test. The exact computational method used to obtain this probability, whether by a closed-form cumulative distribution function, numerical integration, or another approach, is left unspecified, as it does not affect the statistical

logic of the test and varies across distributions.

6.2.1 Tests for Associations

This subsection documents the procedures *GetTest* routes to when the declared variables call for the significance of an association rather than a comparison between groups, for example when two variables are declared as Outputs with no Input, or when a single Input and Output are declared together with no grouping structure. The specific test selected depends on the type of the variables involved: numerical variables route to a correlation test based on the Pearson correlation coefficient, ordinal variables to a rank-based correlation test based on the Spearman coefficient, and categorical variables to a chi-squared test of independence.

Numerical Association: Simple Correlation

Tests whether the linear relationship between one numerical input variable and one numerical output variable is statistically significant [33].

Significance Test

x is the input variable

y is the output variable

n is the sample size; $n_x = n_y$

r is the coefficient calculated using equation (6.18)

$$df = n - 2 \tag{6.31}$$

$$t = r \sqrt{\frac{df}{1 - r^2}} \tag{6.32}$$

$$p = \begin{cases} 2 \cdot P(T_{df} \geq |t|) & \text{two-tailed (testing for any linear relationship)} \\ P(T_{df} \geq t) & \text{one-tailed (testing for a positive relationship)} \\ P(T_{df} \leq t) & \text{one-tailed (testing for a negative relationship)} \end{cases} \tag{6.33}$$

Effect Size: Pearson Correlation Coefficient (r)

The correlation coefficient r serves as the effect size ranging from -1 to 1 where values closer to 0 indicate a weaker effect. By convention, $|0.1|$, $|0.3|$, and $|0.5|$ are considered small, medium, and large effects, respectively [23].

Numerical Association: Multiple Correlation

Multiple correlation measures the strength of the linear relationship between one output variable and multiple input variables in a multiple regression context. It is quantified by the multiple correlation coefficient R , with R^2 representing the proportion of variance in the dependent variable y explained by the independent variables X . In addition to assessing model fit, R^2 serves as an effect size. It shows how much of the variance in the output variable y is explained by the input variables X , providing insight into the practical significance of the model. A higher R^2 indicates a larger effect, while a lower R^2 suggests a weaker relationship. [33, 42, 117]

X is the matrix of input variables, it may also include an intercept; $n \times (k + 1)$

y is the single output variable; $n \times 1$

$\bar{y}$ is the sample mean of y

n is the sample size; number of rows in X and y ; $n = X_{nrow} = y_{nrow}$

m is the number of input variables (and intercept); number of columns in X ; $m = X_{ncol}$

k is the number input variables

df_{model} is the model degrees of freedom; $df_{model} = k$

df_{error} is the error degrees of freedom; $df_{error} = n - m$

df_{total} is the total degrees of freedom; $df_{total} = df_{model} + df_{error}$

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & \cdots & x_{1k} \\ 1 & \vdots & \ddots & \ddots & \vdots \\ 1 & \vdots & \ddots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & \cdots & x_{nk} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} \quad (6.34)$$

Compute the predicted output values:

$$\hat{y} = X \left((X^T X)^{-1} X^T y \right) \quad (6.35)$$

Compute the variation explained by the model:

(i.e. how far the predicted values are from the mean of actual values)

$$SS_{model} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (6.36)$$

$$MS_{model} = \frac{SS_{model}}{df_{model}} \quad (6.37)$$

Compute the unexplained variation:

(i.e. how far actual values deviate from predicted values):

$$SS_{error} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6.38)$$

$$MS_{error} = \frac{SS_{error}}{df_{error}} \quad (6.39)$$

Compute the total variation in the output variable:

(i.e. the overall spread of the actual values around the mean)

$$SS_{total} = SS_{model} + SS_{error} \quad (6.40)$$

$$MS_{total} = \frac{SS_{total}}{df_{total}} \quad (6.41)$$

Significance Test

The significance of the correlation coefficient can be tested using an F -distribution with df_{model} and df_{error} degrees of freedom. The probability value (p) is derived from this distribution.

$$F = \frac{MS_{model}}{MS_{error}} \quad (6.42)$$

$$p = P(F_{df_{model}, df_{error}} \geq F) \quad (6.43)$$

Effect Size: Coefficient of Determination (R^2)

This coefficient measures the proportion of variation in the dependent variable y that can be predicted by the independent variables X . R^2 is a ‘goodness of fit’ measurement that ranges from 0 to 1 and can be interpreted as a percentage. [23]

$$R^2 = \frac{SS_{model}}{SS_{total}} \quad (6.44)$$

Effect Size: Adjusted Coefficient of Determination (R_{adj}^2)

This coefficient measures the proportion of variation in the dependent variable y that can be predicted by the independent variables X adjusted to account for the number of independent variables in the model. R_{adj}^2 ranges from 0 to 1 and can be interpreted as a percentage, it is less than or equal to R^2 . [55]

$$R_{adj}^2 = 1 - \frac{MS_{error}}{MS_{total}} \quad (6.45)$$

Effect Size: Multiple Correlation Coefficient (R)

This coefficient measures the strength (but not the direction) of the correlation. R , or R_{adj} , ranges from 0 to 1. A perfect correlation would be indicated by 1. If R is 0, this would indicate there is no correlation at all. [43]

$$R = \sqrt{R^2} \quad (6.46)$$

$$R_{adj} = \sqrt{R_{adj}^2} \quad (6.47)$$

Ordinal Association: Rank Correlation

Tests whether the monotonic relationship between one ordinal input variable and one ordinal output variable is statistically significant.

Significance Test

x is the input variable

y is the output variable

n is the sample size; $n_x = n_y$

r_s is the coefficient calculated using equation (6.19), (6.22), or (6.23)

$$df = n - 2 \quad (6.48)$$

$$t = r_s \sqrt{\frac{df}{1 - r_s^2}} \quad (6.49)$$

$$p = \begin{cases} 2 \cdot P(T_{df} \geq |t|) & \text{two-tailed (testing for any monotonic relationship)} \\ P(T_{df} \geq t) & \text{one-tailed (testing for a positive relationship)} \\ P(T_{df} \leq t) & \text{one-tailed (testing for a negative relationship)} \end{cases} \quad (6.50)$$

Effect Size: Spearman Correlation Coefficient (r_s)

The correlation coefficient r_s serves as the effect size ranging from -1 to 1 where values closer to 0 indicate a weaker effect. By convention, $|0.1|$, $|0.3|$, and $|0.5|$ are considered small, medium, and large effects, respectively [23].

Categorical Association: Chi-Squared Test of Independence

Tests whether the association between one categorical input variable and one categorical output variable is statistically significant. [33]

Significance Test

x is the input variable, with k categories

y is the output variable, with m categories

n is the total sample size; $n_x = n_y$

The observed counts are arranged into a $k \times m$ contingency table, where each cell O_{ij} is the count of cases falling into category i of x and category j of y :

Table 6.1: Contingency table for the chi-squared test of independence.

	y_1	y_2	$\cdots$	y_m	Total
x_1	O_{11}	O_{12}	$\cdots$	O_{1m}	$n_{1\cdot}$
x_2	O_{21}	O_{22}	$\cdots$	O_{2m}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_k	O_{k1}	O_{k2}	$\cdots$	O_{km}	$n_{k\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot m}$	n

The expected count for each cell, is calculated from the table's row and column totals:

$$E_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n} \quad (6.51)$$

$$df = (k - 1)(m - 1) \quad (6.52)$$

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (6.53)$$

$$p = P(\chi_{df}^2 \geq X^2) \quad (6.54)$$

Bias-corrected (Yates' continuity correction) [43]:

$$\chi_{Yates}^2 = \sum_{i=1}^k \sum_{j=1}^m \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}} \quad (6.55)$$

$$p = P(\chi_{df}^2 \geq \chi_{Yates}^2) \quad (6.56)$$

Effect Size: Cramér's Correlation Coefficient

The correlation coefficient (ϕ obtained from equation (6.24); V obtained from equation (6.26); or $\tilde{V}$ obtained from equation (6.30)) serves as the effect size. ϕ ranges from -1 to 1 where values closer to 0 indicate a weaker effect, and $V/\tilde{V}$ range from 0 to 1, with higher values indicating a stronger association. By convention, $|0.1|$, $|0.3|$, and $|0.5|$ are considered small, medium, and large effects, respectively, though Cohen (1988) provides adjusted, smaller benchmarks as degrees of freedom increase beyond the 2×2 case. [23].

Pairwise Tests

Pairwise association tests extend the corresponding two-variable association test to situations involving more than one input or output variable, or a single categorical input with multiple levels. In the former case, the underlying association test is computed between each relevant pair of variables; in the latter, it is computed separately within each level of the input variable. Unlike the pairwise tests for differences, which may use a fitted approach incorporating model-based corrections (e.g., Tukey’s HSD, Games-Howell), pairwise association tests are comparatively simple: the resulting p -values are either left uncorrected or adjusted using a Bonferroni correction [13, 34]. No other corrections are currently implemented.

p is the unadjusted p -value from the underlying test

k is the number of pairwise comparisons being conducted

$$p_{adj} = p \cdot \binom{k}{2} \quad (6.57)$$

Future Additions

The tests described in this section represent the associations currently implemented within the library, but are not an exhaustive account of all tests for association available in the broader statistical literature. Repeated-measures correlation, canonical correlation analysis, Kendall’s correlation test, and the point-biserial correlation test are well-established tests for association that are not yet implemented and represent possible extensions to this functionality. Additionally, the bias-corrected Cramér’s V coefficient (equation 6.30) is defined within the library but not yet computed in practice, since standard Cramér’s V already incorporates Yates’ correction when calculated from a 2x2 table, making an additional bias correction unlikely to meaningfully change the result in most cases.

6.2.2 Tests For Differences

Categorical Differences

One sample measured once: Chi-Squared Goodness Of Fit

Tests whether the observed frequencies of a single categorical variable are statistically significantly different from a hypothesized distribution

y is the categorical output variable, with m categories

O_i is the observed count in category i

E_i is the expected count in category i under the hypothesized distribution

Significance Test

$$df = m - 1 \quad (6.58)$$

$$\chi^2 = \sum_{i=1}^m \frac{(O_i - E_i)^2}{E_i} \quad (6.59)$$

$$p = P(\chi_{df}^2 \geq \chi^2) \quad (6.60)$$

Effect Size: Cramér's V

Cramér's V (equation 6.26) serves as the effect size. For the goodness-of-fit case, $k = 1$ (a single categorical variable is being tested against an expected distribution), so the denominator reduces to $\min(k - 1, m - 1) = m - 1$ [28].

Two or more samples measured once: Chi-Squared Test of Homogeneity

Tests whether two or more independent samples are drawn from populations with the same distribution of a categorical outcome. This test is computed identically to the chi-squared test of independence presented in Section 6.2.1; the two tests differ only in interpretation, not calculation. A test of independence asks whether two variables are related within a single sample, while a test of homogeneity asks whether several samples have different distributions: the same χ^2 statistic (equation 6.53 or 6.55), degrees of freedom (equation 6.52), and effect size (equation 6.26 or 6.30) apply in both cases [33].

One sample measured twice with two or more possible outcomes: McNemar & McNemar-Bowker Test Of Symmetry [91, 14]

Tests whether the distribution of categorical responses changes significantly between two repeated or paired measurements on the same subjects.

x is the categorical input variable, with two repeated conditions

y is the categorical output variable, with m categories

n is the sample size; $n = n_{x_1} = n_{x_2}$

n_{ij} is the count of subjects whose value of y was category i under the first level of x and category j under the second level of x

The paired observations of y under each level of x are cross-tabulated into an $m \times m$ contingency table:

Significance Test

$$df = \frac{m(m-1)}{2} \quad (6.61)$$

$$X^2 = \sum_{i=1}^{m-1} \sum_{j=i+1}^m \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} \quad (6.62)$$

$$p = P(\chi_{df}^2 \geq X^2) \quad (6.63)$$

Table 6.2: Contingency table for the McNemar and McNemar-Bowker test of symmetry.

		x_2				Total
		y_1	y_2	$\cdots$	y_m	
x_1	y_1	n_{11}	n_{12}	$\cdots$	n_{1m}	$n_{1\cdot}$
	y_2	n_{21}	n_{22}	$\cdots$	n_{2m}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	y_m	n_{m1}	n_{m2}	$\cdots$	n_{mm}	$n_{m\cdot}$
Total		$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot m}$	n

Bias-corrected (Edwards' continuity correction, applicable only to the 2×2 case) [38]:

$$X_{Edwards}^2 = \sum_{i=1}^{m-1} \sum_{j=i+1}^m \frac{(|n_{ij} - n_{ji}| - 1)^2}{n_{ij} + n_{ji}} \quad (6.64)$$

$$p = P(\chi_{df}^2 \geq X_{Edwards}^2) \quad (6.65)$$

Effect Size: Cohen's g

For the $m = 2$ case:

$$g = \frac{\max(n_{12}, n_{21})}{n_{12} + n_{21}} - 0.5 \quad (6.66)$$

For the $m \geq 3$ case:

$$g = \frac{1}{m} \sum_{i=1}^{m-1} \sum_{j=i+1}^m \left(\frac{\max(n_{ij}, n_{ji})}{n_{ij} + n_{ji}} - 0.5 \right) \quad (6.67)$$

Here, Cohen's g indicates, among subjects whose response changed between x_1 and x_2 , whether that change tended to be a genuine shift in one direction or just back-and-forth fluctuation that canceled out. It ranges from 0 to 0.5: values near 0 mean that, for every subject who changed one way (e.g., from "no" to "yes"), roughly another subject changed the opposite way ("yes" to "no"), so the changes offset one another. Values near 0.5 mean nearly all the changes ran in the same direction, indicating a real shift rather than random fluctuation. By convention, g values of 0.05, 0.15, and 0.25 are considered small, medium, and large effects, respectively [23].

One sample measured multiple times with only two possible outcomes: Cochran's Q Test of Differences [21]

Tests whether the proportion of successes differs significantly across three or more repeated or matched binary measurements on the same subjects.

x is the categorical input variable, with k repeated conditions: $x_1, x_2, \dots, x_k$

y is the categorical output variable, with two categories

n is the sample size (number of subjects); $n = n_{x_1} = n_{x_2} = \dots = n_{x_k}$

y_{ij} is the binary-encoded value of y for subject s_i under x_j , coded as 1 for success and 0 for failure

$y_{i\cdot}$ is the total number of successes for subject s_i across all k conditions

$y_{\cdot j}$ is the total number of successes under x_j across all n subjects

p_i is the proportion of successes for subject s_i ; $p_i = \frac{y_{i\cdot}}{k}$

The binary-encoded outcomes of y are arranged into an $n \times k$ table:

Table 6.3: Binary-encoded outcome table for Cochran's Q test of differences.

<i>subject</i>	x_1	x_2	$\dots$	x_k	Total
s_1	y_{11}	y_{12}	$\dots$	y_{1k}	$y_{1\cdot}$
s_2	y_{21}	y_{22}	$\dots$	y_{2k}	$y_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
s_n	y_{n1}	y_{n2}	$\dots$	y_{nk}	$y_{n\cdot}$
Total	$y_{\cdot 1}$	$y_{\cdot 2}$	$\dots$	$y_{\cdot k}$	n

Significance Test

$$df = k - 1 \quad (6.68)$$

$$\chi^2 = \frac{(k-1) \left[k \sum_{j=1}^k y_{\cdot j}^2 - (\sum_{i=1}^n y_{i\cdot})^2 \right]}{k \sum_{i=1}^n y_{i\cdot} - \sum_{i=1}^n y_{i\cdot}^2} \quad (6.69)$$

$$p = P(\chi_{df}^2 \geq \chi^2) \quad (6.70)$$

Effect Size: Chance-Corrected R [10]

$$\delta = \left[k \binom{n}{2} \right]^{-1} \sum_{j=1}^k \sum_{i=1}^{n-1} \sum_{l=i+1}^n |y_{ij} - y_{lj}| \quad (6.71)$$

$$\mu_\delta = \frac{2}{n(n-1)} \left[\left(\sum_{i=1}^n p_i \right) \left(n - \sum_{i=1}^n p_i \right) - \sum_{i=1}^n p_i(1 - p_i) \right] \quad (6.72)$$

Note: the original publication [10] presents the leading coefficient of μ_δ as $\frac{2}{k(k-1)}$; this was later identified as an error and corrected to $\frac{2}{n(n-1)}$ [32].

$$R = 1 - \frac{\delta}{\mu_\delta} \quad (6.73)$$

Chance-corrected R measures how much more consistent subjects' successes were across the k conditions than would be expected if successes were assigned at random. R ranges from 0 to 1, with values closer to 0 indicating that the pattern of successes across conditions was no different than chance, and values closer to 1 indicating that subjects' successes were highly consistent across conditions, well beyond what chance alone would produce.

Pairwise Tests for Categorical Differences

Pairwise tests extend the corresponding omnibus test for categorical differences to identify which specific pairs of groups or conditions differ, following a significant test with three or more samples or conditions. This is accomplished by computing the underlying two-sample test for each relevant pair, with the resulting p -values either left uncorrected or adjusted using a Bonferroni correction [13, 34]. No other corrections are currently implemented.

p is the unadjusted p -value from the underlying two-sample test (6.54, 6.63, or 6.70)

k is the number of pairwise comparisons being conducted

$$p_{adj} = p \cdot \binom{k}{2} \quad (6.74)$$

Future Additions The tests described in this section represent the procedures currently implemented within the library for categorical outcomes, but are not an exhaustive account of all tests for differences available in the broader statistical literature. Fisher's exact test, the G-test of goodness of fit, the G-test of independence, and the binomial test are well-established tests for differences in categorical data that are not yet implemented and represent possible extensions to this functionality. Additionally, the (n -way) Cochran-Mantel-Haenszel test of conditional independence, which extends the chi-squared test of independence to designs stratified by one or more additional categorical factors, is not yet implemented and represents a planned extension for analyzing multi-factor categorical designs.

Ordinal Differences

One sample measured once: Wilcoxon Signed-Ranks Test [147]

Tests whether the median of a single sample is statistically significantly different from a hypothesized median, without assuming the sample is normally distributed.

y is the ordinal output variable

μ_0 is the theoretical comparison value (0 by default)

R_y is the sample sorted by absolute value and ranked

R_y^s is the ranked sample, with a sign applied around μ_0

R_y^+ is the sum of the positive ranks in R_y^s

R_y^- is the sum of the negative ranks in R_y^s

t_j is the j^{th} tie length in R_y

n is the sample size; $n = n_y$

Significance Test

Statistic:

$$W = \min(R_y^+, R_y^-) \quad (6.75)$$

Continuity correction:

$$c = 0.5 \quad (6.76)$$

Tie correction:

$$\sigma = \begin{cases} \sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum(t_j^3 - t_j)}{48}} & \text{(if ties are present)} \\ \sqrt{\frac{n(n+1)(2n+1)}{24}} & \text{(if ties are not present)} \end{cases} \quad (6.77)$$

Normal approximation:

$$Z = \frac{\left| W - \frac{n(n+1)}{4} \right| - c}{\sigma} \quad (6.78)$$

$$p = 2 \cdot P(Z' \geq |Z|) \quad (6.79)$$

Effect Size: Rosenthal's r [41]

$$r = \frac{|Z|}{\sqrt{n}} \quad (6.80)$$

Rosenthal's r serves as the effect size, ranging from 0 to 1, with values closer to 0 indicating a weaker effect and values closer to 1 indicating a stronger effect. By convention, $|0.1|$, $|0.3|$, and $|0.5|$ are considered small, medium, and large effects, respectively [23].

One sample measured twice: Wilcoxon Signed-Ranks Test [147]

Tests whether the median difference between two paired or repeated measurements of the same sample is

statistically significantly different from a hypothesized median, without assuming the differences are normally distributed.

x is the categorical input variable, with two repeated levels: x_1 and x_2

y is the ordinal output variable

y_1 is the value of y under x_1

y_2 is the value of y under x_2

δ is the difference between y_1 and y_2

This test is computed by taking the difference $\delta = y_1 - y_2$ for each paired observation and submitting the resulting sample of differences directly to the one sample Wilcoxon Signed-Ranks Test described above (Equations 6.75–6.80), with y in those equations now taken to be δ , and μ_0 serving as the hypothesized median of the differences. The sample size n is the number of paired observations. No separate equations are needed, since the paired case is mathematically identical to the one sample case.

One sample measured multiple times: Friedman Test [50]

Tests whether the median of an ordinal output variable differs significantly across three or more repeated or matched measurements of the same sample, without assuming the differences are normally distributed.

x is the categorical input variable, with k repeated levels: $x_1, \dots, x_k$

y is the ordinal output variable

n is the number of subjects

nk is the total number of observations

R_{ij} is the rank of subject i 's value of y among the k measurements within that subject's block (i.e., the minimum rank is 1 and the maximum is k)

R_j is the sum of the ranks assigned to the j^{th} measurement, across all subjects

t_j is the j^{th} tie length in the overall ranking

Significance Test [26, 27]

Variance (needed later in post hoc pairwise tests):

$$S^2 = \frac{1}{k-1} \left(\sum_{i=1}^n \sum_{j=1}^k R_{ij}^2 - \frac{nk(k+1)^2}{4} \right) \quad (6.81)$$

Statistic:

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^k \left(R_j - \frac{n(k+1)}{2} \right)^2 \quad (6.82)$$

Tie correction:

$$ties = \sum (t_j^3 - t_j) \quad (6.83)$$

$$Q_t = \frac{Q}{1 - \frac{ties}{nk(k^2-1)}} \quad (6.84)$$

Chi-squared approximation:

$$df = k - 1 \quad (6.85)$$

$$p = P(\chi_{df}^2 \geq Q_t) \quad (6.86)$$

Effect Size: Kendall's W [67, 41]

$$W = \frac{Q_t}{n(k-1)} \quad (6.87)$$

Kendall's W serves as the effect size, ranging from 0 to 1, with values closer to 0 indicating weak agreement in rankings across the k measurements, and values closer to 1 indicating strong agreement.

Two samples measured once: Mann-Whitney U-Test (Wilcoxon Rank-Sum Test) [88, 147]

Tests whether the distribution of an ordinal output variable differs significantly between two independent samples, without assuming the samples are normally distributed.

x is the categorical input variable, with two independent levels: x_1 and x_2

y is the ordinal output variable

R is the ranking of all observations, pooled across both samples

R_{x_1} is the sum of the ranks assigned to the sample under x_1

R_{x_2} is the sum of the ranks assigned to the sample under x_2

t_j is the j^{th} tie length in R

n_{x_1} is the sample size under x_1

n_{x_2} is the sample size under x_2

n is the total sample size; $n = n_{x_1} + n_{x_2}$

Significance Test

Statistic:

$$U_{x_1} = n_{x_1}n_{x_2} + \frac{n_{x_1}(n_{x_1} + 1)}{2} - R_{x_1} \quad (6.88)$$

$$U_{x_2} = n_{x_1}n_{x_2} + \frac{n_{x_2}(n_{x_2} + 1)}{2} - R_{x_2} \quad (6.89)$$

$$U = \min(U_{x_1}, U_{x_2}) \quad (6.90)$$

Continuity correction:

$$c = \begin{cases} 0.5 & \text{if } U < \frac{n_{x_1} n_{x_2}}{2} \\ -0.5 & \text{if } U \geq \frac{n_{x_1} n_{x_2}}{2} \end{cases} \quad (6.91)$$

Tie correction:

$$\sigma = \begin{cases} \sqrt{\frac{n_{x_1} n_{x_2}}{12} \left((n+1) - \frac{\sum (t_j^3 - t_j)}{n(n-1)} \right)} & \text{(if ties are present)} \\ \sqrt{\frac{n_{x_1} n_{x_2} (n+1)}{12}} & \text{(if ties are not present)} \end{cases} \quad (6.92)$$

Normal approximation:

$$Z = \frac{U - \frac{n_{x_1} n_{x_2}}{2} + c}{\sigma} \quad (6.93)$$

$$p = 2 \cdot P(Z' \geq |Z|) \quad (6.94)$$

Effect Size: Rosenthal's r [41]

$$r = \frac{|Z|}{\sqrt{n}} \quad (6.95)$$

Rosenthal's r serves as the effect size, ranging from 0 to 1, with values closer to 0 indicating a weaker effect and values closer to 1 indicating a stronger effect. By convention, $|0.1|$, $|0.3|$, and $|0.5|$ are considered small, medium, and large effects, respectively [23].

Multiple samples measured once: Kruskal-Wallis H Test [75]

Tests whether the distribution of an ordinal output variable differs significantly across three or more independent samples, without assuming the samples are normally distributed.

x is the categorical input variable, with k independent levels: $x_1, \dots, x_k$

y is the ordinal output variable

R is the ranking of all observations, pooled across all samples

R_i is the sum of the ranks assigned to the sample under x_i

R_{ij} is the rank (among all observations) of observation j from the sample under x_i

t_j is the j^{th} tie length in R

n_i is the sample size under x_i

n is the total sample size; $n = \sum_{i=1}^k n_i$

Significance Test [26, 27]

Variance (needed later in post hoc pairwise tests):

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^k \sum_{j=1}^{n_i} R_{ij}^2 - \frac{n(n+1)^2}{4} \right) \quad (6.96)$$

Statistic:

$$H = \left(\frac{12}{n(n+1)} \cdot \sum_{i=1}^k \frac{R_i^2}{n_i} \right) - 3(n+1) \quad (6.97)$$

Tie correction:

$$ties = \sum (t_j^3 - t_j) \quad (6.98)$$

$$H_t = \frac{H}{1 - \frac{ties}{n^3 - n}} \quad (6.99)$$

Chi-squared approximation:

$$df = k - 1 \quad (6.100)$$

$$p = P(\chi_{df}^2 \geq H_t) \quad (6.101)$$

Effect Size: Ordinal Eta-Squared [41]

$$\eta_H^2 = \frac{H_t - k + 1}{n - k} \quad (6.102)$$

Effect Size: Ordinal Epsilon-Squared [41]

$$\varepsilon_R^2 = \frac{H_t(n+1)}{n^2 - 1} \quad (6.103)$$

Both η_H^2 and ε_R^2 serve as effect sizes, ranging from 0 to 1, with higher values indicating a stronger effect. Their interpretation as a proportion of explained variance, by analogy to the parametric eta-squared and epsilon-squared, has been criticized, since the concept of variance does not directly apply to rank-based statistics. By convention, η_H^2 values of 0.01, 0.06, and 0.14 are considered small, medium, and large effects, respectively, following the same thresholds used for ANOVA eta-squared [23]. For ε_R^2 , there is no consensus on classification thresholds; one suggested classification gives small, medium, and large effects at 0.01, 0.08, and 0.26, respectively [41].

Pairwise Tests for Ordinal Differences

Pairwise tests extend the corresponding omnibus test for ordinal differences (Kruskal-Wallis or Friedman) to identify which specific pairs of groups or conditions differ. This class can take two approaches. The first, the *fitted* approach, uses the fitted result of the prior omnibus test to calculate standard errors and corrections for

familywise error; the formal names of tests included in the fitted approach are the Nemenyi-Tukey [97], Dunn-Bonferroni [35], and Conover-Iman-Bonferroni [26] tests. The second, the *unfitted* approach, instead runs appropriate individual two-sample tests and corrects for familywise error using a Bonferroni procedure [13, 34]. No other corrections are currently implemented.

x is the categorical input variable, with k levels/conditions: $x_1, \dots, x_k$

k is the number of samples in the pairwise comparison

n is the number of unique subjects

$\bar{R}_{x_i}$ is the mean rank of the sample under x_i

R_{x_i} is the sum of ranks of the sample under x_i

n_{x_i} is the sample size under x_i

$ties$ is the tie-correction sum from the omnibus test (equation 6.98 or 6.83)

H is the Kruskal-Wallis test statistic from the omnibus test (equation 6.97 or 6.99)

Q is the Friedman test statistic from the omnibus test (equation 6.82 or 6.84)

Fitted Lenient Approach: Nemenyi-Tukey Multiple Comparison Test [97]

Independent samples (after Kruskal-Wallis; apply tie correction):

$$\tau = 1 - \frac{ties}{n^3 - n} \quad (6.104)$$

$$t = \frac{1}{\tau} \cdot \frac{\bar{R}_{x_i} - \bar{R}_{x_j}}{\sqrt{\left(\frac{n(n+1)}{12}\right) \left(\frac{1}{n_{x_i}} + \frac{1}{n_{x_j}}\right)}} \quad (6.105)$$

When ties are not present; Tukey-Kramer approximation:

$$q = |t|\sqrt{2}, \quad df = k \quad (6.106)$$

$$p = P(q'_{df,\infty} \geq q) \quad (6.107)$$

When ties are present; Chi-squared approximation:

$$df = k - 1 \quad (6.108)$$

$$\chi^2 = t^2 \quad (6.109)$$

$$p = P(\chi^2_{df} \geq \chi^2) \quad (6.110)$$

Repeated measures (after Friedman; tie correction does not matter [33]):

$$t = \frac{\bar{R}_{x_i} - \bar{R}_{x_j}}{\sqrt{\frac{k(k+1)}{6n}}} \quad (6.111)$$

$$df = k \quad (6.112)$$

$$q = |t|\sqrt{2} \quad (6.113)$$

$$p = P(q'_{df,\infty} \geq q) \quad (6.114)$$

**Fitted Strict Approach: Dunn Multiple Comparison Test
with Bonferroni Correction [35]**

Uses individual variance estimates that are not pooled across samples/measures.

Independent samples (after Kruskal-Wallis; apply tie correction):

$$\tau = \frac{ties}{12(n-1)} \quad (6.115)$$

$$t = \frac{\bar{R}_{x_i} - \bar{R}_{x_j}}{\sqrt{\left(\frac{n(n+1)}{12} - \tau\right) \left(\frac{1}{n_{x_i}} + \frac{1}{n_{x_j}}\right)}} \quad (6.116)$$

Repeated measures (after Friedman; tie correction does not apply):

$$t = \frac{\bar{R}_{x_i} - \bar{R}_{x_j}}{\sqrt{\frac{k(k+1)}{6n}}} \quad (6.117)$$

For both cases:

$$p = 2 \cdot P(Z' \geq |t|) \quad (6.118)$$

$$p_{adj} = p \cdot \binom{k}{2} \quad (6.119)$$

**Fitted Strict Approach: Conover-Iman Multiple Comparison Test
with Bonferroni Correction [26]**

Uses a shared variance estimate that is pooled across samples/measures.

Independent samples (after Kruskal-Wallis):

$$df = n - k \quad (6.120)$$

$$S^2 = \begin{cases} \text{obtained from equation (6.96)} & \text{(when ties are present)} \\ \frac{n(n+1)}{12} & \text{(when ties are not present)} \end{cases} \quad (6.121)$$

$$t = \frac{\bar{R}_{x_i} - \bar{R}_{x_j}}{\sqrt{S^2 \left(\frac{n-1-H}{df}\right) \left(\frac{1}{n_{x_i}} + \frac{1}{n_{x_j}}\right)}} \quad (6.122)$$

Repeated measures (after Friedman):

$$df = (n - 1)(k - 1) \quad (6.123)$$

$$S^2 = \begin{cases} \text{obtained from equation (6.81)} & \text{(when ties are present)} \\ \frac{kn(k+1)}{12} & \text{(when ties are not present)} \end{cases} \quad (6.124)$$

$$t = \frac{R_{x_i} - R_{x_j}}{\sqrt{S^2 \left(\frac{2n(k-1)}{df} \right) \left(1 - \frac{Q}{n(k-1)} \right)}} \quad (6.125)$$

For both cases:

$$p = 2 \cdot P(T'_{df} \geq |t|) \quad (6.126)$$

$$p_{adj} = p \cdot \binom{k}{2} \quad (6.127)$$

Unfitted Approach: Two-Sample Test with Bonferroni Correction

k is the number of samples in the pairwise comparison

p is the unadjusted probability value calculated in equation (paired 6.79) or (6.94)

$$p_{adj} = p \cdot \binom{k}{2} \quad (6.128)$$

Future Additions

The tests described in this section represent the differences for ordinal outcomes currently implemented within the library, but are not an exhaustive account of all such tests available in the broader statistical literature. The Scheirer-Ray-Hare test [121], a non-parametric extension for factorial designs with ordinal outcomes, is not yet implemented and represents a possible extension to this functionality.

Numerical Differences

One sample measured once: One-Sample T-Test [129]

Tests whether the mean of a single sample is statistically significantly different from a hypothesized population mean, assuming the sample is normally distributed.

y is the numerical output

$\bar{y}$ is the sample mean of y

s_y is the sample standard deviation of y

μ_0 is the hypothesized mean (default is 0)

n is the sample size

Significance Test

$$t = \frac{\bar{y} - \mu_0}{\frac{s_y}{\sqrt{n}}} \quad (6.129)$$

$$df = n - 1 \quad (6.130)$$

$$p = \begin{cases} 2 \cdot P(T_{df} \geq |t|) & \text{two-tailed (testing for any difference)} \\ P(T_{df} \geq t) & \text{one-tailed (testing for a positive difference)} \\ P(T_{df} \leq t) & \text{one-tailed (testing for a negative difference)} \end{cases} \quad (6.131)$$

Effect Size: Cohen's D [23]

This effect size quantifies the strength of the difference between the sample mean and the hypothesized mean in terms of standard deviation units. d ranges from $-\infty$ to ∞ in principle, though in practice it is bounded by the numerical precision of the underlying computation, with values further from 0 indicating a larger effect. By convention, $|0.2|$, $|0.5|$, and $|0.8|$ are considered small, medium, and large effects, respectively [23].

$$d = \frac{\bar{y} - \mu_0}{s_y} \quad (6.132)$$

One sample measured twice: Paired-Sample T-Test [129]

Tests whether the mean difference between two paired or repeated measurements of the same sample is statistically significantly different from a hypothesized mean difference, assuming the differences are normally distributed.

x is the categorical input variable, with two repeated levels: x_1 and x_2

y is the numerical output variable

y_1 is the value of y under x_1

y_2 is the value of y under x_2

δ is the difference between y_1 and y_2

$\bar{\delta}$ is the sample mean of the differences δ

s_δ is the sample standard deviation of the differences δ

μ_0 is the hypothesized mean difference

n is the sample size; $n = n_{y_1} = n_{y_2}$

Significance Test

$$t = \frac{\bar{\delta} - \mu_0}{\frac{s_\delta}{\sqrt{n}}} \quad (6.133)$$

$$df = n - 1 \quad (6.134)$$

$$p = \begin{cases} 2 \cdot P(T_{df} \geq |t|) & \text{two-tailed (testing for any difference)} \\ P(T_{df} \geq t) & \text{one-tailed (testing for a positive difference)} \\ P(T_{df} \leq t) & \text{one-tailed (testing for a negative difference)} \end{cases} \quad (6.135)$$

Effect Size: Cohen's D [23]

This effect size quantifies the strength of the difference between the paired sample mean and the hypothesized mean difference in terms of standard deviation units. By convention, $|0.2|$, $|0.5|$, and $|0.8|$ are considered small, medium, and large effects, respectively [23].

$$d = \frac{\bar{\delta} - \mu_0}{s_\delta} \quad (6.136)$$

Two samples measured once: Two-Sample T-Test [45]

Tests whether the means of two independent samples are statistically significantly different, assuming both samples are normally distributed and have equal variances.

x is the categorical input variable, with two independent levels: x_1 and x_2

y is the numerical output variable

$\bar{y}_{x_1}$ is the sample mean of y under x_1

$\bar{y}_{x_2}$ is the sample mean of y under x_2

$s_{x_1}^2$ is the sample variance of y under x_1

$s_{x_2}^2$ is the sample variance of y under x_2

n_{x_1} is the sample size under x_1

n_{x_2} is the sample size under x_2

n is the total sample size; $n = n_{x_1} + n_{x_2}$

Significance Test

$$s_{pooled} = \sqrt{\frac{s_{x_1}^2(n_{x_1} - 1) + s_{x_2}^2(n_{x_2} - 1)}{n - 2}} \quad (6.137)$$

$$t = \frac{\bar{y}_{x_1} - \bar{y}_{x_2}}{\sqrt{\frac{s_{pooled}^2}{n_{x_1}} + \frac{s_{pooled}^2}{n_{x_2}}}} \quad (6.138)$$

$$df = n - 2 \quad (6.139)$$

$$p = \begin{cases} 2 \cdot P(T_{df} \geq |t|) & \text{two-tailed (testing for any difference)} \\ P(T_{df} \geq t) & \text{one-tailed (testing for a positive difference)} \\ P(T_{df} \leq t) & \text{one-tailed (testing for a negative difference)} \end{cases} \quad (6.140)$$

Effect Size: Cohen's D [23]

This effect size quantifies the strength of the difference between two group means in terms of standard deviation units. By convention, $|0.2|$, $|0.5|$, and $|0.8|$ are considered small, medium, and large effects, respectively [23].

$$d = \frac{\bar{y}_{x_1} - \bar{y}_{x_2}}{s_{pooled}} \quad (6.141)$$

Two samples measured once: Welch's Two-Sample T-Test [144]

Tests whether the means of two independent samples are statistically significantly different, without assuming equal variances between the two samples. Welch's test adjusts the degrees of freedom to account for unequal variances and/or unequal sample sizes, providing a more accurate test than the standard two-sample t-test under these conditions.

x is the categorical input variable, with two independent levels: x_1 and x_2

y is the numerical output variable

$\bar{y}_{x_1}$ is the sample mean of y under x_1

$\bar{y}_{x_2}$ is the sample mean of y under x_2

$s_{x_1}^2$ is the sample variance of y under x_1

$s_{x_2}^2$ is the sample variance of y under x_2

n_{x_1} is the sample size under x_1

n_{x_2} is the sample size under x_2

Significance Test

$$t = \frac{\bar{y}_{x_1} - \bar{y}_{x_2}}{\sqrt{\frac{s_{x_1}^2}{n_{x_1}} + \frac{s_{x_2}^2}{n_{x_2}}}} \quad (6.142)$$

$$df = \frac{\left(\frac{s_{x_1}^2}{n_{x_1}} + \frac{s_{x_2}^2}{n_{x_2}}\right)^2}{\frac{s_{x_1}^4}{n_{x_1}^2(n_{x_1} - 1)} + \frac{s_{x_2}^4}{n_{x_2}^2(n_{x_2} - 1)}} \quad (6.143)$$

$$p = 2 \cdot P(T_{df} \geq |t|) \quad (6.144)$$

Effect Size: Cohen's D [23]

This effect size quantifies the strength of the difference between two group means in terms of standard deviation units, using the average of the two group variances rather than the pooled variance used in the standard two-sample t-test, since equal variances are not assumed here. By convention, $|0.2|$, $|0.5|$, and $|0.8|$

are considered small, medium, and large effects, respectively [23].

$$d = \frac{\bar{y}_{x_1} - \bar{y}_{x_2}}{\sqrt{\frac{s_{x_1}^2 + s_{x_2}^2}{2}}} \quad (6.145)$$

Effect Size: Hedges' G

Hedges' g, a bias-corrected version of Cohen's d appropriate for small or unequal sample sizes. *Note: Not currently implemented:*

$$g = d \cdot \left(1 - \frac{3}{4 \cdot df - 1}\right)$$

Multiple samples measured once: Welch's ANOVA [144, 143]

Tests whether the means of three or more independent samples are statistically significantly different, without assuming equal variances across groups. Welch's ANOVA extends Welch's two-sample t-test to designs with three or more groups, adjusting for differences in variance and sample size to provide a more reliable test than standard ANOVA when group variances differ.

x is the categorical input variable, with k independent levels: $x_1, \dots, x_k$

y is the numerical output variable

$\bar{y}_{x_i}$ is the sample mean of y under x_i

s_{x_i} is the sample standard deviation of y under x_i

n_{x_i} is the sample size under x_i

N is the total sample size across all k groups; $N = \sum_{i=1}^k n_{x_i}$

w_{x_i} is the weight of the sample under x_i ; $w_{x_i} = \frac{n_{x_i}}{s_{x_i}^2}$

w is the total weighted sum; $w = \sum_{i=1}^k w_{x_i}$

Significance Test

$$\bar{y}_{grand}^{Welch} = \frac{\sum_{i=1}^k w_{x_i} \bar{y}_{x_i}}{w} \quad (6.146)$$

$$SS_{effect}^{Welch} = \sum_{i=1}^k w_{x_i} (\bar{y}_{x_i} - \bar{y}_{grand}^{Welch})^2 \quad (6.147)$$

$$\Lambda = \frac{3 \sum_{i=1}^k \frac{(1 - \frac{w_{x_i}}{w})^2}{n_{x_i} - 1}}{k^2 - 1} \quad (6.148)$$

$$df_{effect} = k - 1 \quad (6.149)$$

$$df_{error} = \frac{1}{\Lambda} \quad (6.150)$$

$$F_{Welch} = \frac{\frac{SS_{effect}^{Welch}}{df_{effect}}}{1 + \frac{2\Lambda(k-2)}{3}} \quad (6.151)$$

$$p = P(F_{df_{effect}, df_{error}} \geq F_{Welch}) \quad (6.152)$$

Effect Size: Estimated Omega-Squared [58]

This effect size quantifies the proportion of variance in the dependent variable explained by the independent variable. Omega-squared provides a less biased estimate than eta-squared for smaller samples or when group variances differ [58]. This value ranges from 0 to 1, and can be interpreted similarly to a percentage.

$$\hat{\omega}^2 = \frac{df_{effect}(F_{Welch} - 1)}{df_{effect}(F_{Welch} - 1) + N} \quad (6.153)$$

Multiple samples measured one or more times across one or more outcomes: General Linear Model (includes all ANOVA or MANOVA variants) [96]

The tests in this section are all special cases of the same underlying computation: a design matrix X , built from the declared input variables (including any continuous covariates) and their interactions, is used to predict a response Y , and the variation in Y attributable to a given effect is found by comparing the model's fit with and without that effect included. This comparison, repeated for each effect or interaction in the design, is known as a Type II sum of squares decomposition.

Why did we choose Type II: There are four conventional types of sums of squares decomposition, labeled Type I through Type IV. Type IV addresses a very narrow case of designs and is not discussed further here. Of the remaining three, when a design is balanced (e.g., equal sample sizes), Types I, II, and III all agree, so the choice between them only matters for unbalanced designs. Type I (sequential) SS depends on the order in which factors are entered and is the default in R's base *aov* and *anova* functions. Type III (unique) SS is the most common default among statistical software packages like SAS or SPSS. Type II SS tests meaningful hypotheses for main effects and is found to be the most statistically powerful of the three in unbalanced designs [77] and when no interactions are present [125]. Therefore, like R's *car* package [48], this library defaults to Type II and other decompositions are not currently available.

The same computation applies regardless of whether the input variables are categorical or continuous, whether one outcome or several are declared, and whether observations are independent or repeated on the same subjects; only the construction of X and Y , and the choice of error term against which an effect is tested, depend on these declarations.

X is the design matrix, constructed from the declared input variables (categorical or continuous) and any interactions between them

Y is the response; a single column when one output variable is declared, or a matrix with one column per output variable when multiple outputs are declared

$X^{without}$ is the portion of X that excludes any column associated with the effect being tested, or with any higher-order interaction that includes it; lower-order terms associated with the effect are retained

X^{with} is the portion of X that includes every column except those associated with higher-order interactions that include the effect being tested; this also retains the effect's own columns and any lower-order terms

For a given effect, the model is fit twice: once without the effect, and once with it.

$$\hat{Y}^{without} = Y - X^{without} \left((X^{without})' X^{without} \right)^{-1} (X^{without})' Y \quad (6.154)$$

$$\hat{Y}^{with} = Y - X^{with} \left((X^{with})' X^{with} \right)^{-1} (X^{with})' Y \quad (6.155)$$

The sum of squares (or, when multiple outputs are declared, the sum of squares and cross-products matrix) attributable to the effect is the difference between these two fits:

$$SS^{without} = \left(\hat{Y}^{without} - \text{mean}(\hat{Y}^{without}) \right)' \left(\hat{Y}^{without} - \text{mean}(\hat{Y}^{without}) \right) \quad (6.156)$$

$$SS^{with} = \left(\hat{Y}^{with} - \text{mean}(\hat{Y}^{with}) \right)' \left(\hat{Y}^{with} - \text{mean}(\hat{Y}^{with}) \right) \quad (6.157)$$

$$SS_{effect} = SS^{without} - SS^{with} \quad (6.158)$$

The degrees of freedom for an effect are the product of $(c_j - 1)$ across each input factor j included in that effect, where c_j is the number of levels in factor j (or $c_j = 2$, giving one degree of freedom, when the factor is a continuous covariate):

$$df_{effect} = \prod_j (c_j - 1) \quad (6.159)$$

Once SS_{effect} and df_{effect} are known, the significance of the effect is tested by comparing it against an appropriate error term. When all observations are independent, a single error term, computed from the residuals of the full model, is shared across all effects. When one or more outputs are measured repeatedly on the same subjects, each repeated effect is instead tested against its own error term, derived from that effect's interaction with the subject identifier.

How the input matrix is constructed: The input matrix X , also called a design matrix, is assembled one column group at a time, one group per declared input or interaction. The columns contributed by each type of input are described below, followed by an example showing how they combine.

Numerical inputs

A Numerical input or covariate contributes a single column of its raw values.

Ordinal inputs

An Ordinal input contributes a single column of encoded integers, produced by the *IntegerEncoder* described in Chapter 5.

Categorical inputs

A Categorical input with c levels contributes $c - 1$ columns, produced by the *ReferenceEncoder* described in Chapter 5, one for each non-reference level. For example, a three-level Categorical input (levels A, B, and reference level C) contributes two columns:

Table 6.4: Example encoding of a three-level categorical input using the ReferenceEncoder.

Input	Column: Is A	Column: Is B
A	1	0
B	0	1
C (reference)	0	0

Interactions

A declared interaction contributes the column-wise product of the columns belonging to each of its constituent inputs. For an interaction between two Categorical inputs, every column of the first input's $(c - 1)$ columns is multiplied against every column of the second input's $(d - 1)$ columns, producing $(c - 1) \times (d - 1)$ interaction columns. For example, the interaction between the three-level input above and a second, two-level Categorical input (levels P and reference level Q) produces two interaction columns, formed by multiplying each row's "Is A"/"Is B" values against its "Is P" value:

Table 6.5: Example encoding of an interaction between a three-level and a two-level categorical input.

Cat-Input 1	Cat-Input 2	Is A	Is B	Is P	Is A $\times$ Is P	Is B $\times$ Is P
A	P	1	0	1	1	0
B	P	0	1	1	0	1
A	Q	1	0	0	0	0
C	P	0	0	1	0	0

Assembling X

The resulting sets of columns, one per declared effect (input or interaction), are placed side by side to form X .

Univariate Analysis of Variance

When a single output variable is declared, the computation described in the above reduces to between-subjects, within-subjects, or mixed-design analysis of variance. The sum of squares and degrees of freedom for each effect are calculated as shown in equations (6.158) and (6.159); what remains is selecting the appropriate error term for each effect.

$SS_{error/effect}$ is the sum of squares of the error term used to test a given effect

$df_{error/effect}$ is the degrees of freedom of that error term

When all observations are independent, every effect shares a single error term, computed from the residuals of the full model:

$$MS_{effect} = \frac{SS_{effect}}{df_{effect}} \quad (6.160)$$

$$MS_{error/effect} = \frac{SS_{error/effect}}{df_{error/effect}} \quad (6.161)$$

$$F_{effect} = \frac{MS_{effect}}{MS_{error/effect}} \quad (6.162)$$

Significance Test

$$p = P(F_{df_{effect}, df_{error/effect}} \geq F_{effect}) \quad (6.163)$$

When one or more outputs are measured repeatedly on the same subjects, each within-subjects effect is instead tested against its own error term, derived from that effect's interaction with the subject identifier, rather than the single shared error term used above. Between-subjects effects in such a design continue to use the shared error term. In either case, equations (6.160) through (6.163) are unchanged; only the source of $SS_{error/effect}$ and $df_{error/effect}$ differs.

Effect Size: Partial Eta-Squared

This effect size quantifies the proportion of variance attributable to an effect, relative to that effect and its error term alone, isolating the effect from other unrelated sources of variance in the design. This value ranges from 0 to 1, and can be interpreted similarly to a percentage, though only per effect: with two or more factors in a design, partial eta-squared values across effects are not comparable and do not sum to a meaningful total.

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error/effect}} \quad (6.164)$$

Effect Size: Generalized Eta-Squared [99]

Partial eta-squared uses the error term specific to the effect being tested, which makes it incomparable between a between-subjects design and a repeated-measures design investigating the same factor, since

repeated measurement reduces that error term relative to the between-subjects case. Generalized eta-squared corrects for this by pooling the error terms across all subject-related sources of variance in the design, rather than using only the one error term specific to the effect being tested, making the resulting effect size more directly comparable across designs. This value generally ranges from 0 to 1, and can be interpreted similarly to a percentage.

SS_{pooled_error} is the pooled sum of squares across every source of error associated with subjects in the design; this includes the SS_{error} term itself and $SS_{error/effect_i}$ for every within-subjects effect $i = 1, \dots, w$, where w is the number of within-subjects effects

$$\eta_G^2 = \frac{SS_{effect}}{SS_{effect} + SS_{pooled_error}} \quad (6.165)$$

When all effects in the design are between-subjects, SS_{pooled_error} reduces to $SS_{error/effect}$, and generalized eta-squared is equal to partial eta-squared. As a note, the library does not currently distinguish manipulated from measured factors, so the formula above additionally assumes all factors are manipulated; the full formula presented by Olejnik and Algina (2003) [99] also adjusts for measured factors and blocking variables, which could be a future extension to this library.

Multivariate Analysis of Variance

When two or more output variables are declared, the computation described in the General Framework extends to multivariate analysis of variance.

q is the number of output variables

df_{effect} is the degrees of freedom for the effect (equation 6.159)

$df_{error/effect}$ is the degrees of freedom for the error term used to test the effect

N is the total number of observations

$$\begin{aligned} s &= \min(df_{effect}, q) \\ m_1 &= \frac{|df_{effect} - q| - 1}{2} \\ m_2 &= \frac{df_{error/effect} - q - 1}{2} \end{aligned}$$

As in the univariate case, $SSCP_{effect}$ is obtained from the same computation as equations (6.156) and (6.157); when Y has a single column, the transpose product reduces to the familiar scalar sum of squares, and when Y has multiple columns (i.e. multivariate), it produces the full cross-product matrix instead:

$$SSCP^{without} = \left(\hat{Y}^{without} - \text{mean}(\hat{Y}^{without}) \right)' \left(\hat{Y}^{without} - \text{mean}(\hat{Y}^{without}) \right) \quad (6.166)$$

$$SSCP^{with} = \left(\hat{Y}^{with} - \text{mean}(\hat{Y}^{with}) \right)' \left(\hat{Y}^{with} - \text{mean}(\hat{Y}^{with}) \right) \quad (6.167)$$

$$SSCP_{effect} = SSCP^{without} - SSCP^{with} \quad (6.168)$$

As in the univariate case, the error term $SSCP_{error/effect}$ used to test a given effect is the single shared error term when all effects are between-subjects, or that effect's own interaction with the subject identifier when one or more outputs are measured repeatedly on the same subjects.

Test Statistic: Wilks' Lambda [149]

Wilks' Lambda is the most commonly reported MANOVA test statistic and performs well under most conditions, but is more sensitive than Pillai-Bartlett Trace to violations of the assumption of equal covariance matrices across groups, particularly with small or unequal sample sizes [100].

$$\Lambda = \frac{|SSCP_{error/effect}|}{|SSCP_{effect} + SSCP_{error/effect}|} \quad (6.169)$$

When $df_{effect} = 1$ (two groups), an exact F -statistic is available:

$$F = \frac{1 - \Lambda}{\Lambda} \cdot \frac{df_{error/effect} - q + 1}{q} \quad (6.170)$$

$$df_1 = q \quad (6.171)$$

$$df_2 = df_{error/effect} - q + 1 \quad (6.172)$$

When $df_{effect} > 1$, an approximate F -statistic is used instead:

$$t = \sqrt{\frac{q^2 df_{effect}^2 - 4}{q^2 + df_{effect}^2 - 5}} \quad (6.173)$$

$$w = df_{error/effect} + df_{effect} - \frac{q + df_{effect} + 1}{2} \quad (6.174)$$

$$df_1 = q \cdot df_{effect} \quad (6.175)$$

$$df_2 = wt + 1 - \frac{q \cdot df_{effect}}{2} \quad (6.176)$$

$$F = \frac{(1 - \Lambda^{1/t}) df_2}{\Lambda^{1/t} df_1} \quad (6.177)$$

Test Statistic: Pillai-Barlett Trace [110]

This library reports Wilks' Lambda by default, consistent with its widespread use as the standard MANOVA test statistic in published research. Users seeking greater robustness to violations of the equal covariance assumption, particularly with unbalanced designs, may prefer Pillai-Bartlett Trace instead [100].

$$V = \text{trace} \left((SSCP_{error/effect} + SSCP_{effect})^{-1} SSCP_{effect} \right) \quad (6.178)$$

$$df_1 = s(2m_1 + s + 1) \quad (6.179)$$

$$df_2 = s(2m_2 + s + 1) \quad (6.180)$$

$$F = \frac{(2m_2 + s + 1)V}{(2m_1 + s + 1)(s - V)} \quad (6.181)$$

Test Statistic: Hotelling-Lawley Trace [78, 60]

Hotelling-Lawley Trace can offer greater statistical power than Wilks' Lambda or Pillai-Bartlett Trace when the effect is driven primarily by one outcome variable rather than spread across several, while still accounting for the others in the model. Like Wilks' Lambda, however, it is more sensitive to violations of the equal covariance matrices assumption [100].

$$U = \text{trace} \left(SSCP_{error/effect}^{-1} SSCP_{effect} \right) \quad (6.182)$$

$$df_1 = s \max(df_{effect}, q) \quad (6.183)$$

$$df_2 = 2(sm_2 + 1) \quad (6.184)$$

$$F = \frac{2(sm_2 + 1)U}{s^2(2m_1 + s + 1)} \quad (6.185)$$

Significance Test

For all cases:

$$p = P(F_{df_1, df_2} \geq F) \quad (6.186)$$

Effect Size: Partial Eta-Squared

Partial eta-squared is calculated differently depending on which multivariate test statistic was used, but in each case quantifies the proportion of multivariate variance attributable to the effect, relative to the effect and its error term alone. This value ranges from 0 to 1, and can be interpreted similarly to a percentage.

$$\eta_p^2 = \begin{cases} 1 - \Lambda^{1/s} & \text{(from Wilks' Lambda)} \\ \frac{V}{s} & \text{(from Pillai-Barlett Trace)} \\ \frac{U}{U+s} & \text{(from Hotelling-Lawley Trace)} \end{cases} \quad (6.187)$$

Pairwise Tests for Differences in Numerical Data

Pairwise tests extend the corresponding omnibus test for numerical differences to identify which specific pairs of groups differ, following a significant ANOVA, Welch's ANOVA, or MANOVA with three or more samples. This class can take two approaches. The first, the *fitted* approach, uses the fitted result of the prior omnibus test to calculate standard errors and corrections for familywise error; the formal names of tests included are Tukey's HSD [135], Tukey-Kramer [73], Games-Howell [51], and Estimated Marginal Means [123] with Bonferroni Correction [13, 34]. The second, the *unfitted* approach, instead runs individual two-sample tests and corrects for familywise error using a Bonferroni procedure.

Fitted Lenient Approach: Tukey-Kramer Multiple Comparison Test [135, 73]

Used following a standard ANOVA (equal variances assumed). Tukey's HSD is used when sample sizes are equal; Tukey-Kramer is used otherwise. Because the studentized range distribution already accounts for the number of comparisons being made, no additional correction is applied afterward.

x is the categorical input variable, with k independent levels: $x_1, \dots, x_k$

y is the numerical output variable

$\bar{y}_{x_i}$ is the sample mean of y under x_i

n_{x_i} is the sample size under x_i

$SS_{error/effect}$ and $df_{error/effect}$ are the error sum of squares and degrees of freedom for the effect being analyzed, from the prior omnibus test (equations 6.158 and 6.159)

$df_{correction}$ is the degrees of freedom correction from a significant sphericity assumption test, if applicable; otherwise $df_{correction} = 1$

$$mse = \frac{SS_{error/effect}}{df_{error/effect}} \quad (6.188)$$

Tukey HSD (equal sample sizes):

$$t = \frac{\bar{y}_{x_i} - \bar{y}_{x_j}}{\sqrt{mse \left(\frac{2}{n_{x_i}} \right)}} \quad (6.189)$$

Tukey-Kramer (unequal sample sizes):

$$t = \frac{\bar{y}_{x_i} - \bar{y}_{x_j}}{\sqrt{mse \left(\frac{1}{n_{x_i}} + \frac{1}{n_{x_j}} \right)}} \quad (6.190)$$

In both cases:

$$q = \sqrt{2} |t| \quad (6.191)$$

$$p_{adj} = P \left(q'_{k, df_{error/effect} \cdot df_{correction}} \geq q \right) \quad (6.192)$$

Fitted Lenient Approach: Games-Howell Multiple Comparison Test [51]

Used following Welch's ANOVA (equal variances not assumed). As with Tukey HSD, no additional correction is applied beyond the studentized range distribution.

x is the categorical input variable, with k independent levels: $x_1, \dots, x_k$

y is the numerical output variable

$\bar{y}_{x_i}$ is the sample mean of y under x_i

$s_{x_i}^2$ is the sample variance of y under x_i

n_{x_i} is the sample size under x_i

$$t = \frac{\bar{y}_{x_i} - \bar{y}_{x_j}}{\sqrt{\frac{s_{x_i}^2}{n_{x_i}} + \frac{s_{x_j}^2}{n_{x_j}}}} \quad (6.193)$$

$$df_{Welch} = \frac{\left(\frac{s_{x_i}^2}{n_{x_i}} + \frac{s_{x_j}^2}{n_{x_j}} \right)^2}{\frac{s_{x_i}^4}{n_{x_i}^2(n_{x_i} - 1)} + \frac{s_{x_j}^4}{n_{x_j}^2(n_{x_j} - 1)}} \quad (6.194)$$

$$q = \sqrt{2} |t| \quad (6.195)$$

$$p_{adj} = P(q'_{k, df_{Welch}} \geq q) \quad (6.196)$$

Fitted Strict Approach: Estimated Marginal Means with Bonferroni Correction [123]

Used following a standard ANOVA. Unlike Tukey HSD, this test uses the t -distribution directly and requires an explicit Bonferroni correction afterward.

x is the categorical input variable, with k independent levels: $x_1, \dots, x_k$

y is the numerical output variable

$\bar{y}_{x_i}$ is the sample mean of y under x_i

n_{x_i} is the sample size under x_i

$SS_{error/effect}$ and $df_{error/effect}$ are as defined above

$df_{correction}$ is as defined above

$$mse = \frac{SS_{error/effect}}{df_{error/effect}} \quad (6.197)$$

$$t = \frac{\bar{y}_{x_i} - \bar{y}_{x_j}}{\sqrt{mse \left(\frac{1}{n_{x_i}} + \frac{1}{n_{x_j}} \right)}} \quad (6.198)$$

$$p = 2 \cdot P(T'_{df_{error/effect} \cdot df_{correction}} \geq |t|) \quad (6.199)$$

$$p_{adj} = p \cdot \binom{k}{2} \quad (6.200)$$

Unfitted Approach: Two-Sample Test with Bonferroni Correction and Pooled Standard Deviation

Used when variances are assumed equal and samples are independent. The pooled standard deviation is computed across all k groups in the comparison, not only the pair being tested.

x is the categorical input variable, with k independent levels: $x_1, \dots, x_k$

y is the numerical output variable

$\bar{y}_{x_i}$ is the sample mean of y under x_i

$s_{x_i}^2$ is the sample variance of y under x_i

n_{x_i} is the sample size under x_i

$$s_{pooled} = \sqrt{\frac{\sum_{i=1}^k s_{x_i}^2 (n_{x_i} - 1)}{\sum_{i=1}^k (n_{x_i} - 1)}} \quad (6.201)$$

$$df_{pooled} = \sum_{i=1}^k (n_{x_i} - 1) \quad (6.202)$$

$$t = \frac{\bar{y}_{x_i} - \bar{y}_{x_j}}{s_{pooled} \sqrt{\frac{1}{n_{x_i}} + \frac{1}{n_{x_j}}}} \quad (6.203)$$

$$p = 2 \cdot P(T'_{df_{pooled}} \geq |t|) \quad (6.204)$$

$$p_{adj} = p \cdot \binom{k}{2} \quad (6.205)$$

Two-Sample Test with Bonferroni Correction

Used for paired, Welch's, or independent two-sample comparisons as appropriate, depending on the design of the omnibus test being followed up.

k is the number of samples in the pairwise comparison

t is the test statistic calculated in equation (6.133), (6.142), or (6.138)

df is the degrees of freedom calculated in equation (6.134), (6.143), or (6.139)

$df_{correction}$ is as defined above

$$p = 2 \cdot P(T'_{df \cdot df_{correction}} \geq |t|) \quad (6.206)$$

$$p_{adj} = p \cdot \binom{k}{2} \quad (6.207)$$

Future Additions

The tests described in this section represent the differences for numerical outcomes currently implemented within the library, but are not an exhaustive account of all such tests available in the broader statistical literature. Several extensions to the General Linear Model framework are not yet implemented, including analysis of covariance (ANCOVA), multivariate analysis of covariance (MANCOVA), and their repeated-measures and mixed-design counterparts; and a multivariate one-sample and paired-sample tests (Hotelling's T^2). Several alternative corrections for familywise error in pairwise post hoc comparisons; such as the Benjamini-Hochberg, Holm-Bonferroni, Sidák, and Scheffé procedures are possible extensions beyond the Bonferroni correction currently supported.

6.2.3 Assumption Tests

[43] In addition to the tests for differences and associations described above, the library implements several assumption tests that can be used alongside a given test for differences to verify that the conditions underlying a chosen statistical procedure are met by the data. Levene's test and the Brown-Forsythe test assess the assumption of equal variances across independent groups. Mauchly's test assesses the assumption of sphericity in repeated-measures and mixed-design analyses, with the Greenhouse-Geisser and Huynh-Feldt corrections available when this assumption is violated. Box's M test assesses the assumption of equal covariance matrices across groups in multivariate designs (all implemented under *CompareVariances*). Bartlett's test of sphericity (*CheckReducibility*) and the Kaiser-Meyer-Olkin (KMO) measure (*CheckReducibilityStrength*) assess the suitability of a correlation matrix for procedures such as factor analysis. The Shapiro-Wilk test assesses the assumption that a sample is normally distributed (*CompareDistributionToNormal*). These assumption tests are not described in detail here, but are noted as a part of the library's broader testing infrastructure, available to support and verify the procedures described throughout this chapter.

6.2.4 Routing Logic

The preceding sections of this chapter describe, one procedure at a time, the mathematics underlying each test this library implements. This section describes the other half of the system: how the type and role declarations a user makes are used to automatically select the correct procedure from among them, without requiring the user to know its name in advance. This routing logic is the library's central contribution. As described in Study 1 and Study 2, users consistently struggled not with performing a statistical test once one was identified, but with knowing which test applied to their data in the first place; the routing logic documented here is what allows a user to make declarations about their data, rather than decisions about statistical procedures, and have the correct procedure selected on their behalf.

The procedure returned by *GetTest* depends on the number and type of Output variables and the number, type, and role of Input variables. Tables 6.6–6.11 give the complete routing logic, organized first by Output structure and then by Input count, type, and role. For example, a two-level categorical Input paired with

a numerical Output routes to an independent samples t -test under the Input role, but a paired t -test under the Repeated role.

Not every combination has a corresponding procedure implemented; these are marked *not yet supported* rather than omitted, so the full space of valid designs remains visible. Unsupported procedures are concentrated in more involved cases, including logistic and ordinal regression and multivariate extensions such as MANCOVA and Hotelling's t^2 -test.

Table 6.6: GetTest routing: no output declared

Inputs	Input	Type(s)	Input	Role(s)	Procedure
1	Any	Any			<i>Error: No test available</i>
2+	Numerical	Input			PCA/EFA
2+	Ordinal	Any			<i>rank-based PCA: not yet supported</i>
2+	Categorical	Any			<i>MCA: not yet supported</i>
2+	Mixed types	Any			<i>FAMD: not yet supported</i>

Table 6.7: GetTest routing: one numerical output

Inputs	Input	Type(s)	Input	Role(s)	Procedure
0	—	—			One-sample t -test
1	Numerical	Input			Pearson correlation
2+	Numerical	Input			Multiple regression
1	Ordinal	Input			Spearman correlation
2+	Ordinal	Input			Multiple regression
1	Categorical (2 levels)	Input			Independent-samples t -test
1	Categorical (2 levels)	Repeated			Paired t -test
1	Categorical (3+ levels)	Input			One-way ANOVA
1	Categorical (3+ levels)	Repeated			One-way repeated-measures ANOVA
2+	Categorical	Input			n -way ANOVA
2+	Categorical	Repeated			Repeated-measures factorial ANOVA
2+	Categorical	Repeated			Mixed-design ANOVA
2+	Mixed	Mixed			<i>ANCOVA: not yet supported</i>

Table 6.8: GetTest routing: multiple numerical outputs

Inputs	Input Type	Input Role(s)	Procedure
0	—	—	Pearson correlation
0	—	Identifier	<i>Repeated-measures corr.: not yet supported</i>
1	Numerical	Input	Pearson correlation (input vs. each output)
2+	Numerical	Input	Multivariate multiple regression
1	Categorical (2 levels)	Input	<i>Hotelling's t^2-test: not yet supported</i>
1	Categorical (2 levels)	Repeated	<i>Hotelling's Paired t^2-test: not yet supported</i>
1	Categorical (3+ levels)	Input	One-way MANOVA
1	Categorical (3+ levels)	Repeated	One-way repeated-measures MANOVA
2+	Categorical	Input	n -way MANOVA
2+	Categorical	Repeated	Mixed-design MANOVA
2+	Mixed	Mixed	<i>MANCOVA: not yet supported</i>

Table 6.9: GetTest routing: one ordinal output

Inputs	Input Type	Input Role(s)	Procedure
0	—	—	One-sample Wilcoxon signed-rank test
1	Numerical or Ordinal	Input	Spearman correlation
2+	Numerical or Ordinal	Input	<i>Ordinal regression: not yet supported</i>
1	Categorical (2 levels)	Input	Mann-Whitney U test
1	Categorical (2 levels)	Repeated	Paired Wilcoxon signed-rank test
1	Categorical (3+ levels)	Input	Kruskal-Wallis test
1	Categorical (3+ levels)	Repeated	Friedman test
2+	Categorical	Input	<i>Ordinal regression: not yet supported</i>

Table 6.10: GetTest routing: one categorical output

Inputs	Input Type	Input Role(s)	Procedure
0	—	—	Chi-square goodness-of-fit test
0	Categorical (2 levels)	Repeated	McNemar/McNemar-Bowker test
0	Categorical (3+ levels)	Repeated	Cochran's Q test
1+	Numerical and/or Ordinal	Input	<i>Logistic regression: not yet supported</i>
1	Categorical (2+ levels)	Input	Chi-square test of homogeneity
2+	Categorical	Input	<i>Logistic regression: not yet supported</i>
2+	Mixed	Input	<i>Logistic regression: not yet supported</i>

Table 6.11: GetTest routing: no input, multiple outputs (pairwise correlation matrix)

Output Types	Procedure
Numerical v. Numerical	Pearson correlation
Ordinal v. Ordinal	Spearman correlation
Numerical v. Ordinal	Spearman correlation
Categorical v. Categorical	Cramér's V correlation
Categorical v. Numerical	<i>NA: not yet supported</i>
Categorical v. Ordinal	<i>NA: not yet supported</i>

6.3 Predictive Modeling

GetModel (naming to be determined) is the library's entry point for predictive modeling. Where *GetTest* answers inferential questions about existing data, *GetModel* is used when the analytical goal is to build a model capable of generating predictions for future observations. As with *GetSummary* and *GetTest*, the user declares the type and role of each variable, and the system resolves the appropriate model, encoding, and preprocessing automatically. *GetModel* returns an *UntrainedModel* object, which can be fit to the declared data via a *Train()* call to produce a *TrainedModel*, capable of generating predictions on new data via *Predict()*.

```
DataFrame frame
frame:Load("data.csv")
CategoricalVariable color = frame:CreateCategoricalInputVariable("color")
NumericalVariable age = frame:CreateNumericalInputVariable("age")
OrdinalVariable rating = frame:CreateOrdinalInputVariable("rating")
NumericalVariable price = frame:CreateNumericalOutputVariable("price")
ModelDesign design
design:AddVariable(color)
design:AddVariable(age)
design:AddVariable(rating)
design:AddVariable(price)
UntrainedModel model = frame:GetModel(design)
TrainedModel trained = model:Train()
```

Figure 6.3: Experimental interface pseudocode for generating an inferential test.

6.3.1 Supervised Learning

One or more numerical outputs: Linear Regression

Linear regression is selected by *GetModel* whenever one or more numerical output variables are declared, regardless of the type of the declared input variables. Input variables of any type (numerical, ordinal, or categorical) are encoded into the design matrix X as part of the model's preprocessing, prior to fitting. The library currently implements linear regression via a single closed-form ordinary least squares (OLS) solution, computed identically whether one or several output variables are declared.

X is the $n \times m$ matrix of encoded input variables, which may include a column for the intercept

Y is the $n \times q$ matrix of output variables, where q is the number of declared output variables ($q = 1$ for a single output)

W is the $m \times q$ matrix of fitted coefficients (weights), with one column per output variable

$$W = (X^T X)^{-1} X^T Y \quad (6.208)$$

Each column of W contains the coefficients for one output variable, estimated jointly with all other declared outputs in a single matrix operation. Once fitted, predictions for new observations are generated by applying these coefficients to a newly encoded input matrix X_{new} :

$$\hat{Y}_{new} = X_{new} W \quad (6.209)$$

The encoding of categorical and ordinal input variables into X (and X_{new}) follows the declarative process described in Chapter 5.

6.3.2 Routing Logic

GetModel's routing logic follows the same declarative principle established for *GetTest* in Section 6.2.4: the type and role of each declared variable determine which procedure is selected, without requiring the user to name a specific model. At the highest level, the presence or absence of a declared output variable distinguishes supervised from unsupervised procedures; among supervised procedures, the type of the declared output variable narrows the selection further. A numerical output currently routes to linear regression, with categorical and ordinal outputs intended to route to logistic and ordinal logistic regression, respectively, once implemented.

This routing logic is necessarily incomplete, reflecting the limited scope of predictive procedures implemented thus far. Unlike *GetTest*'s routing logic, which is documented here in full table form, *GetModel*'s routing is described narratively due to the smaller number of procedures currently implemented. A more complete system would need to account for other considerations, such as overlap between fundamentally different model families (e.g., tree-based vs. gradient-descent-based models) that is not always reducible to type and role alone, more complex architectures such as multi-layer neural networks, hyperparameter selection, and decisions around data splitting and randomization. Establishing sensible, usable defaults for these considerations, consistent with the declarative philosophy underlying the rest of this library, is identified as a direction for future work.

6.4 Scope of This Chapter

This chapter has aimed to document the mathematical foundations of the library's major statistical and predictive procedures. It does not articulate an exhaustive account of every procedure currently implemented within the library. Some implemented functionality (e.g., principal component analysis, k-means clustering, matrix decomposition, alternative effect sizes, analysis corrections and assumption tests) falls outside of the inferential and predictive procedures detailed here. The intent of this chapter is to demonstrate the depth, accuracy, and extensibility of the library's mathematical foundations through representative coverage, rather than to serve as a complete reference for every feature within. Despite the effort made to provide accurate depictions of all procedures discussed here, it is likely that some errors remain, given the scope and complexity of the material. Readers are encouraged to independently verify any equation of particular interest before relying on it.

Chapter 7

Conclusion

My journey into the uncomfortable side of statistics and data science did not start as an investigation of statistics anxiety. As an undergraduate, I was brought onto the Quorum team to help build out a charting library that was in its infancy. Later, as a graduate student, I was hired on a grant specifically to explore whether changes to statistical software, particularly its output, could reduce statistical anxiety. I was tasked with building out a statistics library based on the new procedure names from our previous study [72], and with developing a 1000-person anxiety study alongside it. If you have read this far, you know I never reached one thousand participants, but the work continued regardless. Building the library meant learning the statistics properly, and it meant confronting, quickly and repeatedly, how inconsistent the tools, terminology, and even the interpretations of the same test could be across sources.

Anxiety turned out to be harder to capture than expected. Existing scales ran up to fifty items and did not mention tools or software at all. Some items were too easy, asking about procedures like t-tests or correlations that did not expose much of a problem, while other procedures assumed a level of statistical background well beyond what a large sample of non-statisticians would have. The literature pointed me toward self-efficacy as a complementary construct, and pairing the two gave a much clearer picture of what participants were actually experiencing. Somewhere in designing that first study, I also decided not to stop at output alone: if I was asking participants to read statistical results, it made sense to also ask them to select the right test and write the code for it.

That decision produced a result I did not expect. The simple data-typing questions, recognizing whether a variable was numerical, categorical, and so on, turned out to be where participants struggled the least in Study 1. That was the moment the real idea emerged: if people already reason well about their data without needing to memorize procedure names or operations, why build an interface that requires both? A use-what-you-already-know architecture, one built around the judgments people already make correctly, seemed like the more promising direction, and it became the basis for the declarative interface at the center of Study 2.

Study 1 used free-form text code, which, while workable, left too much room for participants to trip over unrelated syntax rather than the statistical concepts I actually cared about. Around the same time, our lab was making progress on block-based programming, and the idea of using blocks to remove that syntactic noise stuck with me. Study 1 was also limited to simple procedures, t-tests and one-way ANOVAs, and the effects, while present, were not large. Since Study 2 no longer required reading raw output, I could push the tasks toward genuinely complex designs (repeated measures, mixed ANOVAs, MANOVA) without burying participants in jargon they would otherwise have had to parse. Participants performed better with that complexity than I expected, and, more surprisingly, were far more optimistic about it than I expected as well.

Taken together, the two studies point to something more specific than “design reduces anxiety.” The clearest finding, and the one I think matters most, came from Study 2: control-group participants who hit a wall they could not get past did not report the anxiety their situation would predict. Their anxiety scores looked statistically indistinguishable from participants who succeeded, but their confidence collapsed. Struggling silently, in other words, does not necessarily register as struggle on a self-report anxiety scale; it registers instead as a form of disengagement that anxiety measures alone would miss entirely. If this pattern holds outside of this study, it has a real implication for how statistics education interventions get evaluated: anxiety scores can look acceptable while participants are actually disengaging.

The declarative layer and Quorum Statistics Library, described in Chapters 5 and 6, are not a full re-implementation of everything tested in Study 2, and I think that distinction is worth stating directly rather than leaving it to be inferred from the chapter ordering. What they demonstrate is that the type-and-role vocabulary the studies found effective, numerical, ordinal, categorical, input, output, repeated, is expressive enough to specify real statistical and predictive procedures end to end, with mathematics that holds up to scrutiny. That scrutiny proved worthwhile independent of its original purpose; verifying these procedures by hand surfaced a real inconsistency in a widely used R package along the way. Finishing the library to the point where it can run Study 2’s actual tasks, and testing it in a real class setting, is the natural next step.

7.1 Downstream Opportunities

The contributions described in this dissertation (typed variable declarations, semantic role assignment, and automated test routing) were motivated primarily by usability and anxiety reduction. However, the same information that makes a statistical interface more navigable for human users also makes several downstream tasks more tractable for the system itself. The remainder of this chapter outlines three areas where the typed data science framework established here opens concrete research opportunities.

Structured visualization. Producing an appropriate visualization of a statistical result currently requires either user specification or heuristics applied to raw data. When the system knows that an input variable is categorical with three ordered levels, that an output is continuous, and that the analysis is a one-way ANOVA

with a pairwise follow-up, it has enough information to generate a well-formed figure (appropriate axis types, correct uncertainty representation, meaningful labeling) without needing additional information. This is not the same as natural-language chart generation, which must infer structure from data values. Because the typed framework makes structure explicit before analysis begins, chart generation becomes a deterministic mapping from a known type signature rather than an inference problem. A natural direction for future work is to study whether automatically generated, semantically grounded visualizations improve interpretability for novice users relative to standard plotting defaults, and whether the reduction in specification burden affects engagement in the same way that typed test selection did in this work.

Plain-language output generation. Study 1 in this dissertation found that output interpretation was the area where users struggled most and where design changes produced the largest gains. The experimental interface addressed this by pre-generating structured, sentence-form output with APA-formatted values. A typed framework could extend this further: when the system knows the variable types, the test family, whether covariates are present, and whether the design is repeated or between-subjects, it can generate output narratives that are not just readable but *accurate* with respect to the structure of the analysis. The risk in plain-language statistical reporting is that simplification introduces ambiguity (e.g., saying “the groups differed” without specifying on which variable and what it means in regards to the other variables). Type and role information constrains that ambiguity directly, because the system knows what was compared, how, and at what level of the design. Investigating whether this approach reduces misinterpretation relative to either raw software output or hand-written summaries is a tractable and practically motivated research question.

Interpretable machine learning. A declarative framework has direct implications for the interpretability of machine learning models. One persistent challenge in applied ML is that encoding decisions made upstream (i.e. one-hot or dummy encoding of categorical variables or ordinal encoding) fracture the feature space in ways that downstream interpretation tools do not automatically reverse. A SHAP value [82], for example, is computed per encoded column, meaning a categorical variable with five levels produces five separate values, none of which corresponds directly to the feature a user understands as meaningful. When the system retains the semantic identity of each variable across the encoding step, it can aggregate value back to the original feature level, presenting a single interpretable contribution per conceptual variable rather than per encoded fragment. Similarly, coefficient interpretation in regression is complicated by encoding choices that are rarely made explicit to the user: a coefficient on a dummy-coded categorical variable is interpretable only relative to the reference category, which many users do not know was chosen or what it is. A declarative framework in which the reference level is declared explicitly, and in which the system generates output that names the comparison being made, addresses this problem at the interface level rather than requiring the user to reconstruct it from documentation. These benefits would be direct consequences of having variable semantics available at evaluation and interpretation time, and they represent a promising direction for future work at

the intersection of declarative statistical computing and human-centered machine learning.

7.2 Closing

None of this dissertation asks whether something new can be built. Typed variables, declarative test selection, plain-language output, these are not new ideas so much as a willingness to ask whether decisions made decades ago still deserve to be the default. Statistical software has largely inherited its conventions rather than reconsidered them, and much of what feels difficult about learning statistics may have less to do with the statistics themselves than with vocabulary and interface choices that have not been revisited in a long time. This work does not resolve that question so much as show it is worth asking seriously, and that asking it produces measurable, reproducible differences in how people experience learning a subject that has quietly intimidated people for a very long time.

Appendix A

Study 1

A.1 Tasks

This appendix presents the complete set of figures depicting the exact tasks and questions as presented to participants in the study. For each of the six tasks, two figures are provided: one showing the Traditional interface and one showing the Experimental interface. These visuals allow readers to directly compare design features, layout, and wording across conditions, providing context for the performance, confidence, and anxiety results reported in the main text. Figures include all instructions, prompts, and question formats that participants encountered during the study.

Task 1 of 6: Selection

You are given a dataset called **apples.csv**, where each row corresponds to a single apple and each column corresponds to a variable. Your goal is to determine the types of variables and the best statistical tests to understand the data.

Below is a snippet of **apples.csv**:

apple	weight	tree height	color	taste
Apple 1	154.5	20.6	red	sweet
Apple 2	150.1	20.5	green	tart
**	**	**	**	**

** the rows have been truncated, assume there are more than listed and assume there is no missing data.

The column labeled **apple** is the apple's unique identification.
The column labeled **weight** is the apple's weight in grams.
The column labeled **tree height** is the height of the tree in feet that grew the apple.
The column labeled **color** is the apple's color (red, green or yellow).
The column labeled **taste** is the apple's taste (sweet, tart or both).

Question 1: What type of variable is weight?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 2: What type of variable is tree height?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 3: What type of variable is color?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 4: What type of variable is taste?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 5: In the first row of the dataset, were the values for weight, tree height, color and taste collected from the same apple?

- ☐ Yes
- ☐ No
- ☐ I am not sure
- ☐ There is not enough information

Question 6: Which statistical test would be most appropriate to measure the strength of the relationship between an apple's weight and its tree height?

- ☐ CorrelateGroups()
- ☐ CompareMeans()
- ☐ CompareVariances()
- ☐ CompareCounts()
- ☐ I am not sure
- ☐ There is not enough information

Question 7: Which statistical test would be most appropriate to measure if there is a relationship between an apple's color and its taste?

- ☐ CorrelateGroups()
- ☐ CompareMeans()
- ☐ CompareVariances()
- ☐ CompareCounts()
- ☐ I am not sure
- ☒ There is not enough information

Figure A.1: Task 1 (Experimental version): Selection. Correct answers – weight and tree height: quantitative; color and taste: qualitative; first row same apple: yes; test for weight vs. tree height: *CorrelateGroups()*; test for color vs. taste: *CompareCounts()*.

Task 1 of 6: Selection

You are given a dataset called **apples.csv**, where each row corresponds to a single apple and each column corresponds to a variable. Your goal is to determine the types of variables and the best statistical tests to understand the data.

Below is a snippet of **apples.csv**:

apple	weight	tree height	color	taste
Apple 1	154.5	20.6	red	sweet
Apple 2	150.1	20.5	green	tart
**	**	**	**	**

** the rows have been truncated, assume there are more than listed and assume there is no missing data.

The column labeled **apple** is the apple's unique identification.
The column labeled **weight** is the apple's weight in grams.
The column labeled **tree height** is the height of the tree in feet that grew the apple.
The column labeled **color** is the apple's color (red, green or yellow).
The column labeled **taste** is the apple's taste (sweet, tart or both).

Question 1: What type of variable is weight?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 2: What type of variable is tree height?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 3: What type of variable is color?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 4: What type of variable is taste?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 5: In the first row of the dataset, were the values for weight, tree height, color and taste collected from the same apple?

- ☐ Yes
- ☐ No
- ☐ I am not sure
- ☐ There is not enough information

Question 6: Which statistical test would be most appropriate to measure the strength of the relationship between an apple's weight and its tree height?

- ☐ cor.test()
- ☐ aov()
- ☐ t.test()
- ☐ chisq.test()
- ☐ I am not sure
- ☐ There is not enough information

Question 7: Which statistical test would be most appropriate to measure if there is a relationship between an apple's color and its taste?

- ☐ cor.test()
- ☐ aov()
- ☐ t.test()
- ☐ chisq.test()
- ☐ I am not sure
- ☐ There is not enough information

Figure A.2: Task 1 (Traditional version): Selection. Correct answers – weight and tree height: quantitative; color and taste: qualitative; first row same apple: yes; test for weight vs. tree height: `cor.test()`; test for color vs. taste: Traditional = `chisq.test()`.

Task 2 of 6: Selection

You are given a dataset called **heartrates.csv**, where each row corresponds to a single animal and each column corresponds to a variable. Your goal is to determine the types of variables and the best statistical tests to understand the data.

Below is a snippet of **heartrates.csv**:

id	species	bpm
1	lion	56
2	lion	63
3	lion	55
4	bear	58
5	goat	72
6	horse	70
7	bear	56
**	**	**

** the rows have been truncated, assume there are more than listed and assume there is no missing data.

The column labeled **id** is a unique identification for the animal in each row.
The column labeled **species** is one of four species groups (lion, bear, goat and horse).
The column labeled **bpm** is the beats per minute heartrate measured for that animal.

Question 1: What type of variable is species?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 2: What type of variable is bpm?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 3: What best describes the species variable?

- ☐ It is an independent, input, explanatory or predictor variable.
- ☐ It is a dependent, output, explained or measured variable.
- ☐ I am not sure
- ☐ There is not enough information

Question 4: What best describes the bpm variable?

- ☐ It is an independent, input, explanatory or predictor variable.
- ☐ It is a dependent, output, explained or measured variable.
- ☐ I am not sure
- ☐ There is not enough information

Question 5: Was the heartrate (bpm) measured for each individual animal more than once?

- ☐ Yes
- ☐ No
- ☐ I am not sure
- ☐ There is not enough information

Question 6: Which statistical test would be most appropriate to determine if the average bpm is different between the four species?

- ☐ CorrelateGroups()
- ☐ CompareMeans()
- ☐ CompareVariances()
- ☐ CompareCounts()
- ☐ I am not sure
- ☐ There is not enough information

Figure A.3: Task 2 (Experimental version): Selection. Correct answers – species: qualitative; bpm: quantitative; species as predictor, bpm as outcome; repeated measurements: no; test for comparing bpm across species: *CompareMeans()*.

Task 2 of 6: Selection

You are given a dataset called **heartrates.csv**, where each row corresponds to a single animal and each column corresponds to a variable. Your goal is to determine the types of variables and the best statistical tests to understand the data.

Below is a snippet of **heartrates.csv**:

id	species	bpm
1	lion	56
2	lion	63
3	lion	55
4	bear	58
5	goat	72
6	horse	70
7	bear	56
**	**	**

** the rows have been truncated, assume there are more than listed and assume there is no missing data.

The column labeled **id** is a unique identification for the animal in each row.
The column labeled **species** is one of four species groups (lion, bear, goat and horse).
The column labeled **bpm** is the beats per minute heartrate measured for that animal.

Question 1: What type of variable is species?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 2: What type of variable is bpm?

- ☐ Quantitative, Numerical, Continuous, or Interval
- ☐ Qualitative, Factor, Categorical, or Nominal
- ☐ Ordinal, Ordered, Ranked or Hierarchical
- ☐ I am not sure
- ☐ There is not enough information

Question 3: What best describes the species variable?

- ☐ It is an independent, input, explanatory or predictor variable.
- ☐ It is a dependent, output, explained or measured variable.
- ☐ I am not sure
- ☐ There is not enough information

Question 4: What best describes the bpm variable?

- ☐ It is an independent, input, explanatory or predictor variable.
- ☐ It is a dependent, output, explained or measured variable.
- ☐ I am not sure
- ☐ There is not enough information

Question 5: Was the heartrate (bpm) measured for each individual animal more than once?

- ☐ Yes
- ☐ No
- ☐ I am not sure
- ☐ There is not enough information

Question 6: Which statistical test would be most appropriate to determine if the average bpm is different between the four species?

- ☐ cor.test()
- ☐ aov()
- ☐ t.test()
- ☐ chisq.test()
- ☐ I am not sure
- ☐ There is not enough information

Figure A.4: Task 2 (Traditional version): Selection. Correct answers – species: qualitative; bpm: quantitative; species as predictor, bpm as outcome; repeated measurements: no; test for comparing bpm across species: `aov()`.

Task 3 of 6: Input

You are given an example of how to write the code for a statistical test using `example.csv`. Your goal is to write, adapt or modify the code from the example for the new dataset `heartrates.csv` to determine if there is a difference in **bpm** between the four **species** groups. If you are using a screen reader, please turn on all-punctuation verbose mode when examining the example code or writing the code for this portion of the survey.

JAWS: Insert+V
 NVDA: Insert+N
 VoiceOver: Control+Option+V

Below is a snippet of the example dataset `example.csv` and the example code:

groups	values
A	13.2
B	12.5
**	**

** the rows have been truncated, assume there are more than listed.

```
DataFrame frame
frame:Load("example.csv")
frame:SelectInputFactor("groups")
frame:SelectOutputVariable("values")
frame:TestAssumptions(true)
frame:CompareMeans()
```

Below is a snippet of `heartrates.csv`:

id	species	bpm
1	lion	56
2	lion	63
3	lion	55
4	bear	58
5	goat	72
6	horse	70
7	bear	56
**	**	**

** the rows have been truncated, assume there are more than listed.

The column labeled **id** is a unique identification for the animal in each row.
 The column labeled **species** is one of four species groups (lion, bear, goat and horse).
 The column labeled **bpm** is the beats per minute heartrate measured for that animal.

Enter your code here

Type 'unsure' if you are not sure what to do to complete this task.

Type 'not enough information' if there is not enough information provided to complete this task.

Figure A.5: Task 3 (Experimental version): Input. Correct answers – file name: `"heartrates.csv"`; input factor: `"species"`; output variable: `"bpm"`; main test: `CompareMeans()`.

Task 3 of 6: Input

You are given an example of how to write the code for a statistical test using `example.csv`. Your goal is to write, adapt or modify the code from the example for the new dataset `heartrates.csv` to determine if there is a difference in **bpm** between the four **species** groups. If you are using a screen reader, please turn on all-punctuation verbose mode when examining the example code or writing the code for this portion of the survey.

JAWS: Insert+V
NVDA: Insert+N
VoiceOver: Control+Option+V

Below is a snippet of the example dataset `example.csv` and the example code:

groups	values
A	13.2
B	12.5
**	**

** the rows have been truncated, assume there are more than listed.

```
frame = read.csv(file="example.csv")
frame$groups = as.factor(frame$groups)
aov_result = aov(values~groups, data=frame)
shapiro.test(aov_result$residuals)
leveneTest(aov_result)
summary(aov_result)
effectsize::eta_squared(aov_result)
```

Below is a snippet of `heartrates.csv`:

id	species	bpm
1	lion	56
2	lion	63
3	lion	55
4	bear	58
5	goat	72
6	horse	70
7	bear	56
**	**	**

** the rows have been truncated, assume there are more than listed.

The column labeled **id** is a unique identification for the animal in each row.
The column labeled **species** is one of four species groups (lion, bear, goat and horse).
The column labeled **bpm** is the beats per minute heartrate measured for that animal.

Enter your code here

Type 'unsure' if you are not sure what to do to complete this task.

Type 'not enough information' if there is not enough information provided to complete this task.

Figure A.6: Task 3 (Traditional version): Input. Correct answers – file name: `"heartrates.csv"`; input factor: `as.factor(frame$species)`; output variable: `bpm`; main test: `aov(bpm~species, data=frame)`.

Task 4 of 6: Input

You are given an example of how to write the code for a statistical test using [example.csv](#).

Your goal is to write, adapt or modify the code from the example for the new dataset **medication.csv** to determine if heart rates are affected by a trial of a medication, and also if there is a difference between the two groups.

If you are using a screen reader, please turn on all-punctuation verbose mode when examining the example code or writing the code for this portion of the survey.

JAWS: Insert+V

NVDA: Insert+N

VoiceOver: Control+Option+V

Below is a snippet of the example dataset [example.csv](#) and the example code:

id	group	measure1	measure2	measure3
1	A	4.2	2.6	5.1
2	B	3.7	3.1	2.8
**	**	**	**	**

** the rows have been truncated, assume there are more than listed.

```
DataFrame frame
frame:Load("example.csv")
frame:SelectInputFactor("group")
frame:SelectRepeatedFactor("measures", "measure1,measure2,measure3")
frame:CompareMeans()
```

Below is a snippet of **medication.csv**:

person	group	before	during	after
1	group1	56	60	54
2	group2	70	75	70
3	group2	65	72	62
4	group2	72	82	70
5	group1	78	88	74
**	**	**	**	**

** the rows have been truncated, assume there are more than listed.

The column labeled **person** is a unique identification for the person in each row.
The column labeled **group** is one of two groups (group1 or group2).
The column labeled **before** is the resting heartrate before taking the medication.
The column labeled **during** is the resting heartrate while taking the medication.
The column labeled **after** is the resting heartrate after completing the medication.

Enter your code here

Type 'unsure' if you are not sure what to do to complete this task.

Type 'not enough information' if there is not enough information provided to complete this task.

Figure A.7: Task 4 (Experimental version): Input. Correct answers – file name: *"medication.csv"*; repeated factor variable: any name in first parameter of *frame:SetRepeatedFactor("...", "...")*; repeated factor levels: *"after,during,before"* in the second parameter of *frame:SetRepeatedFactor("...", "...")*; output variable: generated automatically from *frame:SetRepeatedFactor("...", "...")*; main test: *CompareMeans()*. See Psychometrics (Section 3.4.10) for full breakdown of grading criteria.

Task 4 of 6: Input

You are given an example of how to write the code for a statistical test using [example.csv](#).

Your goal is to write, adapt or modify the code from the example for the new dataset **medication.csv** to determine if heart rates are affected by a trial of a medication, and also if there is a difference between the two groups.

If you are using a screen reader, please turn on all-punctuation verbose mode when examining the example code or writing the code for this portion of the survey.

JAWS: Insert+V
NVDA: Insert+N
VoiceOver: Control+Option+V

Below is a snippet of the example dataset [example.csv](#) and the example code:

id	group	measure1	measure2	measure3
1	A	4.2	2.6	5.1
2	B	3.7	3.1	2.8
**	**	**	**	**

** the rows have been truncated, assume there are more than listed.

```
frame = read.csv(file="example.csv")
long_frame = gather(frame, key = "measures", value =
"responses", "measure1", "measure2", "measure3")
long_frame$id = as.factor(long_frame$id)
long_frame$group = as.factor(long_frame$group)
long_frame$measures = as.factor(long_frame$measures)
aov_result = aov(responses~group*measures + Error(id/measures),
data = long_frame)
summary(aov_result)
effectsize::eta_squared(aov_result)
```

Below is a snippet of **medication.csv**:

person	group	before	during	after
1	group1	56	60	54
2	group2	70	75	70
3	group2	65	72	62
4	group2	72	82	70
5	group1	78	88	74
**	**	**	**	**

** the rows have been truncated, assume there are more than listed.

The column labeled **person** is a unique identification for the person in each row.
The column labeled **group** is one of two groups (group1 or group2).
The column labeled **before** is the resting heartrate before taking the medication.
The column labeled **during** is the resting heartrate while taking the medication.
The column labeled **after** is the resting heartrate after completing the medication.

Enter your code here

Type 'unsure' if you are not sure what to do to complete this task.

Type 'not enough information' if there is not enough information provided to complete this task.

Figure A.8: Task 4 (Traditional version): Input. Correct answers – file name: *"medication.csv"*; repeated factor variable: any consistent name set in the second parameter (*key*) of *gather(...)*; repeated factor levels: *"after"*, *"during"*, *"before"* in the fourth, fifth and sixth parameters of *gather(...)*; output variable: any consistent name set in the third parameter (*value*) of *gather(...)*; subject Identifier: *as.factor(long_frame\$person)*; main test: should use all consistent variable and object names in *aov(...)*. See Psychometrics (Section 3.4.10) for full breakdown of grading criteria.

Task 5 of 6: Output

You are given the output for a statistical test.

Your goal is to interpret the output and answer a few questions concerning it.

If you are using a screen reader, please turn on all-punctuation verbose mode when examining the code for this portion of the survey.

JAWS: Insert+V

NVDA: Insert+N

VoiceOver: Control+Option+V

Output to interpret

Summary

A comparison between 2 samples was tested. First, the samples were tested for normality and all samples appear to be normal. Results show that there may not be a difference between group1 and group2, $t(58.99) = 1.61$, $p = .112$, Cohen's $d = 0.41$.

Table Summary

comparison	statistic (t)	degrees of freedom (df)	probability (p)	significant (p < .05)	effect size (cohen's d)	effect size magnitude
group1_group2	1.61	58.99	0.112	no	0.41	small

Descriptive Table Summary

sample	size (n)	mean (μ)	standard deviation (s)
group1	31	11.62	3.53
group2	30	10.18	3.46

Input code

```
DataFrame frame
frame:Load("data.csv")
frame:SelectInputFactor("groups", "group1,group2")
frame:TestDistributionAssumption(true)
frame:CompareMeans()
```

The input code is simply for your reference, you do not need to interpret this.

Question 1: This test assumes that both samples are approximately normally distributed, do the samples appear to meet this requirement?

☐ Yes

☐ No

☐ I am not sure

☐ There is not enough information

Question 2: This test is able to handle unequal sample sizes, but are group1 and group2 the same size?

☐ Yes

☐ No

☐ I am not sure

☐ There is not enough information

Question 3: Would the means of group1 and group2 be considered different from each other at a significance level of 0.05?

☐ Yes

☐ No

☐ I am not sure

☐ There is not enough information

Question 4: What is the strength of the effect size of the observed interpretation between group1 and group2?

☐ Small

☐ Medium

☐ Large

☐ None

☐ I am not sure

☐ There is not enough information

Figure A.9: Task 5 (Experimental version): Output. Correct answers – normality assumption: yes; sample sizes equal: no; statistical significance: no; Cohen's d: small.

Task 5 of 6: Output

You are given the output for a statistical test.

Your goal is to interpret the output and answer a few questions concerning it.

If you are using a screen reader, please turn on all-punctuation verbose mode when examining any code for this portion of the survey.

JAWS: Insert+V

NVDA: Insert+N

VoiceOver: Control+Option+V

Output to interpret

```

Shapiro-Wilk normality test
data: group1
W = 0.968, p-value = 0.4659
-----
Shapiro-Wilk normality test
data: group2
W = 0.97853, p-value = 0.7852
-----
Welch Two Sample t-test data:
group1 and group2
t = 1.6112, df = 58.99, p-value = 0.1125
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
-0.3490012 3.2341625
sample estimates:
mean of x mean of y
11.62258 10.18000
-----
Cohen's d | 95% CI
-----
0.41 | [-0.10, 0.92]

- Estimated using pooled SD.Warning message:
Missing values detected. NAs dropped.
-----
group1 group2
11.62258 10.18000
-----
group1 group2
3.531450 3.461254
-----
group1 group2
31 30
-----

```

Input code

```

frame = read.csv(file="data.csv")
group1 = frame$group1
group2 = frame$group2
shapiro.test(group1)
shapiro.test(group2)
t.test(group1, group2)
effectsize::cohens_d(group1, group2)
sapply(frame, mean, na.rm = T)
sapply(frame, sd, na.rm = T)
colSums(!is.na(frame))

```

The input code is simply for your reference, you do not need to interpret this.

Question 1: This test assumes that both samples are approximately normally distributed, do the samples appear to meet this requirement?

☐ Yes

☐ No

☐ I am not sure

☐ There is not enough information

Question 3: Would the means of group1 and group2 be considered different from each other at a significance level of 0.05?

☐ Yes

☐ No

☐ I am not sure

☐ There is not enough information

Question 2: This test is able to handle unequal sample sizes, but are group1 and group2 the same size?

☐ Yes

☐ No

☐ I am not sure

☐ There is not enough information

Question 4: What is the strength of the effect size of the observed interpretation between group1 and group2?

☐ Small

☐ Medium

☐ Large

☐ None

☐ I am not sure

☐ There is not enough information

Figure A.10: Task 5 (Traditional version): Output. Correct answers – normality assumption: yes; sample sizes equal: no; statistical significance: no; Cohen’s d: small.

Task 6 of 6: Output

You are given the output for a statistical test.

Your goal is to interpret the output and answer a few questions concerning it.

If you are using a screen reader, please turn on all-punctuation verbose mode when examining any code for this portion of the survey.

JAWS: Insert+V

NVDA: Insert+N

VoiceOver: Control+Option+V

Output to interpret

Summary							
A comparison to measure the effect of Group on Response was tested using 4 samples. First, the residuals were tested for normality and all residuals appear to be normal. Second, the variances were tested and appear to be similar between the 4 samples. Results show that it is likely there is a difference in the samples under the effect Group ($F(3, 118) = 3.11, p = .029$, eta-squared = .073). Follow up pairwise comparisons for the effect Group show that group1 likely differs from group3 ($p = .015$). However, all other pairs under the effect Group appear similar.							
Table Summary							
comparison	statistic (F)	degrees of freedom (dfn)	error degrees of freedom (dfd)	probability (p)	significant (p < .05)	effect size (eta-squared)	effect size magnitude
Group	3.113	3	118	0.029	yes	0.073	medium
Pairwise Table Summary							
pairwise (p < 0.05)	group1	group2	group3	group4			
group1			0.015				
group2							
group3	0.015						
group4							
Descriptive Table Summary							
sample	size (n)	mean (μ)	standard deviation (σ)				
group1	31	11.62	3.53				
group2	30	10.27	3.77				
group3	31	8.94	3.04				
group4	30	10.18	3.46				

Question 1: This test assumes that all the samples have equal variances, do the samples appear to meet this requirement?

- ☐ Yes
☐ No
☐ I am not sure
☐ There is not enough information

Question 2: This test assumes that the residuals of the data are approximately normally distributed, do the residuals appear to meet this requirement?

- ☐ Yes
☐ No
☐ I am not sure
☐ There is not enough information

Question 3: Would the effect of Group be considered significant at significance level of 0.05?

- ☐ Yes
☐ No
☐ I am not sure
☐ There is not enough information

Input code

```

DataFrame frame
frame:Load("data.csv")
frame:SelectInputFactor("Group")
frame:SelectOutputVariable("Response")
frame:TestAllAssumptions(true)
frame:CompareMeans()
  
```

The input code is simply for your reference, you do not need to interpret this.

Question 4: What is the strength of the observed interpretation of the effect Group if any?

- ☐ Small
☐ Medium
☐ Large
☐ None
☐ I am not sure
☐ There is not enough information

Question 5: Which pair of samples are significantly different if any?

- ☐ group1 vs group2
☐ group1 vs group3
☐ group1 vs group4
☐ group2 vs group3
☐ group2 vs group4
☐ group3 vs group4
☐ No sample pairs appear different
☐ I am not sure
☐ There is not enough information

Figure A.11: Task 6 (Experimental version): Output. Correct answers – homogeneity of variance: yes; normality: yes; group effect significant: yes; effect size (eta-squared): medium; Tukey post hoc: group1 vs. group3.

Task 6 of 6: Output

You are given the output for a statistical test.

Your goal is to interpret the output and answer a few questions concerning it.

If you are using a screen reader, please turn on all-punctuation verbose mode when examining any code for this portion of the survey.

JAWS: Insert+V
NVDA: Insert+N
VoiceOver: Control+Option+V

Output to interpret

```
Shapiro-Wilk normality test

data: aov_result$residuals
W = 0.98775, p-value = 0.3446

-----
Levene's Test for Homogeneity of Variance (center = median)
Df F value Pr(>F)
group 3 0.2283 0.8765
118
-----
Anova Table (Type II tests)
Response: Response
Sum Sq Df F value Pr(>F)
Group 111.6 3 3.113 0.02895 *
Residuals 1410.1 118
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

-----
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = Response ~ Group, data = frame)
$Group
diff lwr upr p adj
group2-group1 -1.35591398 -3.663101 0.9512730 0.4220520
group3-group1 -2.68064516 -4.968843 -0.3924477 0.0146748
group4-group1 -1.44258065 -3.749768 0.8646064 0.3661811
group3-group2 -1.32473118 -3.631918 0.9824558 0.4429681
group4-group2 -0.08666667 -2.412688 2.2393548 0.9996703
group4-group3 1.23806452 -1.069122 3.5452515 0.5028554
-----
Group
0.07333938

-----
# A tibble: 4 x 4
Group mean sd size
<fct> <dbl> <dbl> <int>
1 group1 11.6 3.53 31
2 group2 10.3 3.77 30
3 group3 8.94 3.04 31
4 group4 10.2 3.46 30
-----
```

Input code

```
frame = read.csv(file="data.csv")
frame$Group = as.factor(frame$Group)
aov_result = aov(Response~Group, data = frame)
shapiro.test(aov_result$residuals)
leveneTest(aov_result)
Anova(aov_result, type=2)
TukeyHSD(aov_result)
eta_squared(aov_result)
group_by(frame, Group) %>% summarise(mean = mean(Response), sd
= sd(Response) , size= n())
```

The input code is simply for your reference, you do not need to interpret this.

Question 1: This test assumes that all the samples have equal variances, do the samples appear to meet this requirement?

- ☐ Yes
- ☐ No
- ☐ I am not sure
- ☐ There is not enough information

Question 2: This test assumes that the residuals of the data are approximately normally distributed, do the residuals appear to meet this requirement?

- ☐ Yes
- ☐ No
- ☐ I am not sure
- ☐ There is not enough information

Question 3: Would the effect of Group be considered significant at significance level of 0.05?

- ☐ Yes
- ☐ No
- ☐ I am not sure
- ☐ There is not enough information

Question 4: What is the strength of the observed interpretation of the effect Group if any?

- ☐ Small
- ☐ Medium
- ☐ Large
- ☐ None
- ☐ I am not sure
- ☐ There is not enough information

Question 5: Which pair of samples are significantly different if any?

- ☐ group1 vs group2
- ☐ group1 vs group3
- ☐ group1 vs group4
- ☐ group2 vs group3
- ☐ group2 vs group4
- ☐ group3 vs group4
- ☐ No sample pairs appear different
- ☐ I am not sure
- ☐ There is not enough information

Figure A.12: Task 6 (Traditional version): Output. Correct answers – homogeneity of variance: yes; normality: yes; group effect significant: yes; effect size (eta-squared): medium; Tukey post hoc: group1 vs. group3.

A.2 Statistical Analysis Result Tables

For the sake of transparency and completeness, this appendix contains all statistical tables that were not included in the main text. While the body of the paper highlights the most relevant results, the full set of analyses is presented here to allow readers to examine the data in greater detail. Tables that were not central to the narrative have been moved here to streamline the main article while still providing complete access to all findings.

A.2.1 Measure Validation

Table A.1: Independent grading agreement between Rater A and Rater B for a subset of code responses. The values shown were used to calculate Cohen’s kappa reported in the main text, providing a measure of inter-rater reliability.

		Rater A		
		1	0	Total
Rater B	1	639	14	653
	0	17	185	202
	Total	656	199	855

Table A.2: Test-retest correlations between participants' first and second task ratings in each area. Higher correlations indicate greater consistency of participants' responses across similar tasks, providing evidence of the reliability of these measures. Overall, the results suggest reasonable reliability for both anxiety and confidence ratings in all areas.

Task Area	N	Measure	p-value	Correlation (r)
Selection (T1 vs. T2)	283	Anxiety	< 0.001 ***	0.70
		Confidence	< 0.001 ***	0.67
Input (T3 vs. T4)	271	Anxiety	< 0.001 ***	0.82
		Confidence	< 0.001 ***	0.80
Output (T5 vs. T6)	266	Anxiety	< 0.001 ***	0.82
		Confidence	< 0.001 ***	0.78

A.2.2 Study-Wide

Table A.3: Correlation between participants' average anxiety and confidence ratings across the study, task areas and aspects. Consistent with prior research, anxiety and confidence were moderately negatively correlated, supporting the validity of the measurement system.

Task Area	Aspect Scale	p-value	<i>r</i>
Study-Wide	Overall	< 0.001 ***	-0.29
	Overall	< 0.001 ***	-0.36
	The format of the data	< 0.001 ***	-0.38
	The explanation of the data	< 0.001 ***	-0.43
Selection	Determining the types of variables	< 0.001 ***	-0.40
	Determining the most appropriate statistical test	< 0.001 ***	-0.36
	The complexity of the questions	< 0.001 ***	-0.42
	The naming convention of the statistical tests	< 0.001 ***	-0.36
	Familiarity with the terms or concepts	< 0.001 ***	-0.37
	Overall	< 0.001 ***	-0.32
	The format of the data	< 0.001 ***	-0.34
	The explanation of the data	< 0.001 ***	-0.33
	The format of the code	< 0.001 ***	-0.32
Input	The explanation of the task	< 0.001 ***	-0.38
	The example code	< 0.001 ***	-0.35
	The amount of code	< 0.001 ***	-0.36
	The naming convention of the code	< 0.001 ***	-0.33
	The complexity of the code	< 0.001 ***	-0.34
	Familiarity with the code, terms or concepts	< 0.001 ***	-0.32
	Overall	< 0.001 ***	-0.30
	The format of the output	< 0.001 ***	-0.30
	The amount of output	< 0.001 ***	-0.27
	The complexity of the output	< 0.001 ***	-0.27
Output	The numerical values of the output	< 0.001 ***	-0.26
	The tables (if any) in the output	< 0.001 ***	-0.25
	The complexity of the questions	< 0.001 ***	-0.22
	The code used to produce the output	0.021 *	-0.14
	Familiarity with the terms or concepts	< 0.001 ***	-0.29

Table A.4: ANOVA results for anxiety and confidence on neutral data- and information-related items. No significant main effects of design were observed. The design $\times$ gender interaction showed females in the Traditional group were slightly more anxious and less confident than males; controlling for academic field reduced this effect.

Source	Anxiety					Confidence				
	SS	df	F	p	η_p^2	SS	df	F	p	η_p^2
Design	0.80	1	0.81	.369	.003	0.62	1	0.34	.559	.001
Gender	7.91	1	8.03	.005**	.028	3.93	1	2.18	.141	.008
Design $\times$ Gender	3.93	1	3.99	.047*	.014	7.14	1	3.96	.048*	.014
Residuals	274.86	279				503.60	279			

Table A.5: ANOVA results for performance on neutral data- and format-related questions. No significant main effects of design were observed. The design $\times$ gender interaction showed females in the Traditional group were slightly more uncertain than males; controlling for academic field reduced this effect.

Source	Uncertainty					Correctness				
	SS	df	F	p	η_p^2	SS	df	F	p	η_p^2
Design	103	1	0.40	.526	.001	251	1	1.00	.318	.004
Gender	137	1	0.54	.464	.002	37	1	0.15	.701	.001
Design $\times$ Gender	1242	1	4.86	.028*	.017	164	1	0.66	.419	.002
Residuals	71308	279				69333	277			

A.2.3 Selection Area

Table A.6: Post hoc Tukey–Kramer tests for study-wide analyses. Significant pairwise differences follow up the ANCOVA results reported in the main text. This indicates that participants who took at least one statistics course generally showed higher correctness and lower uncertainty and those who took multiple felt more confident as well. Gender differences in confidence were often related to academic field, participants in the formal sciences were most confident, and females in other fields were significantly less confident than all other participants. Post hoc comparisons are shown only for significant effects that involve more than two groups; effects with only two groups are reported in the main text.

Comparison	p-value
<i>Statistics Course Experience (Confidence)</i>	
None ($M = 2.92, SD = 1.20$) < Multiple ($M = 3.62, SD = 1.38$)	0.0009
One ($M = 3.24, SD = 1.06$) < Multiple ($M = 3.62, SD = 1.38$)	0.0163
(Other comparisons were not significant)	
<i>Statistics Course Experience (Uncertainty)</i>	
None ($M = 44.3, SD = 26.7$) > One ($M = 28.9, SD = 22.0$)	0.0003
None ($M = 44.3, SD = 26.7$) > Multiple ($M = 22.3, SD = 19.8$)	< 0.0001
(Other comparisons were not significant)	
<i>Statistics Course Experience (Correctness)</i>	
None ($M = 55.4, SD = 16.4$) < One ($M = 63.9, SD = 16.0$)	0.0230
None ($M = 55.4, SD = 16.4$) < Multiple ($M = 67.4, SD = 15.9$)	0.0002
(Other comparisons were not significant)	
<i>Gender $\times$ Academic Field (Confidence)</i>	
Female Formal Sciences ($M = 3.39, SD = 1.24$) > Female Other ($M = 2.71, SD = 1.18$)	0.0238
Male Formal Sciences ($M = 3.28, SD = 1.11$) > Female Other ($M = 2.71, SD = 1.18$)	0.0013
Male Other ($M = 3.46, SD = 1.28$) > Female Other ($M = 2.71, SD = 1.18$)	0.0016
(Other comparisons were not significant)	

Table A.7: ANCOVA results for selection area uncertainty (proportion of “not sure” responses across all questions) and correctness (calculated after excluding “not sure” responses). The design reduced uncertainty, indicating participants were more decisive when selecting statistical tests. Statistics experience also predicted lower uncertainty, yet not higher correctness, suggesting increased decisiveness did not equate to greater accuracy. Gender influenced uncertainty (men were more uncertain) and academic field influenced accuracy (formal sciences were more correct).

Source	Uncertainty					Correctness				
	SS	df	F	p	η_p^2	SS	df	F	p	η_p^2
Design	6005	1	26.19	< 0.001***	0.090	769	1	3.36	0.068	0.010
Gender	1550	1	6.76	0.010**	0.020	187	1	0.81	0.367	0.000
Academic Field	800	1	3.49	0.063	0.010	1128	1	4.93	0.027*	0.020
Primary Lang	40	1	0.17	0.678	0.000	214	1	0.93	0.335	0.000
Age Group	477	1	2.08	0.150	0.000	161	1	0.70	0.403	0.000
Has Disability	116	1	0.51	0.477	0.000	133	1	0.58	0.446	0.000
Stats Course Exp	5979	2	13.04	< 0.001***	0.090	707	2	1.54	0.216	0.010
Anxiety	21	1	0.09	0.764	0.000	480	1	2.10	0.149	0.000
Confidence	3791	1	16.54	< 0.001***	0.060	901	1	3.94	0.048*	0.010
Design $\times$ Gender	7	1	0.03	0.866	0.000	18	1	0.08	0.780	0.000
Gender $\times$ Field	229	1	1.00	0.319	0.000	44	1	0.19	0.660	0.000
Residuals	61905	270				61360	268			

A.2.4 Input Area

Table A.8: Selection area ANCOVA results for overall and nomenclature-related ratings of anxiety and confidence. Design did not significantly affect participants' overall anxiety or confidence, but it did yield a positive effect on confidence when ratings were limited to nomenclature-related items (i.e., subs-scales reflecting participants' reactions to statistical test names during task selection). This suggests that while overall perceptions were similar across groups, the experimental design improved participants' confidence specifically in navigating statistical terminology. Several effects emerged for covariates (e.g., disability status, statistics course experience, age).

Source	Anxiety					Confidence				
	SS	df	F	p	η_p^2	SS	df	F	p	η_p^2
<i>Overall Ratings</i>										
Design	0.002	1	0.002	0.967	0.000	1.550	1	1.108	0.293	0.004
Gender	3.655	1	3.495	0.063	0.012	14.60	1	10.46	0.001**	0.036
Academic Field	0.060	1	0.057	0.811	0.000	0.550	1	0.395	0.530	0.001
Primary Lang	0.819	1	0.783	0.377	0.003	0.820	1	0.591	0.443	0.002
Age Group	0.631	1	0.603	0.438	0.002	4.600	1	3.296	0.070	0.010
Has Disability	10.936	1	10.46	0.001**	0.036	5.380	1	3.850	0.051	0.013
Stats Courses Exp	5.784	2	2.765	0.065	0.019	17.65	2	6.319	0.002**	0.043
Design $\times$ Gender	0.240	1	0.229	0.632	0.001	15.97	1	11.43	< 0.001***	0.039
Gender $\times$ Field	0.149	1	0.142	0.707	0.000	4.620	1	3.309	0.070	0.010
Residuals	294.95	282				393.88	282			
<i>Nomenclature-Related Ratings</i>										
Design	1.580	1	1.073	0.301	0.004	12.54	1	6.783	0.010**	0.023
Gender	0.850	1	0.578	0.448	0.002	2.10	1	1.134	0.288	0.004
Academic Field	2.570	1	1.739	0.188	0.006	0.76	1	0.409	0.523	0.001
Primary Lang	0.330	1	0.222	0.638	0.001	3.67	1	1.983	0.160	0.007
Age Group	1.660	1	1.122	0.290	0.004	20.34	1	11.00	0.001**	0.038
Has Disability	19.21	1	13.01	0.001***	0.044	0.98	1	0.531	0.467	0.002
Stats Courses Exp	15.08	2	5.105	0.007**	0.035	25.78	2	6.972	0.001**	0.047
Design $\times$ Gender	3.260	1	2.206	0.139	0.008	1.660	1	0.898	0.344	0.003
Gender $\times$ Field	0.000	1	0.000	0.993	0.000	2.210	1	1.198	0.275	0.004
Residuals	416.45	282				521.32	282			

Table A.9: Tukey–Kramer tests for selection area. The gender gap in confidence did not persist when considering only nomenclature-related items, suggesting more of a data-related difference. Having taken more than one statistics course consistently helped promote higher confidence in the selection area, and having taken at least one course helped promote engagement and decisiveness.

Comparison	p-value
<i>Design $\times$ Gender (Confidence: Overall Ratings)</i>	
Traditional Male ($M = 4.38, SD = 1.16$) > Traditional Female ($M = 3.34, SD = 1.17$)	0.0001
Traditional Male ($M = 4.38, SD = 1.16$) > Experimental Male ($M = 3.87, SD = 1.26$)	0.0289
(Other comparisons were not significant)	
<i>Statistics Course Experience (Confidence: Overall Ratings)</i>	
None ($M = 3.68, SD = 1.26$) < Multiple ($M = 4.34, SD = 1.35$)	0.0013
One ($M = 4.01, SD = 1.16$) < Multiple ($M = 4.34, SD = 1.35$)	0.0596
(Other comparisons were not significant)	
<i>Statistics Course Experience (Uncertainty)</i>	
None ($M = 19.4, SD = 19.0$) > One ($M = 10.7, SD = 12.6$)	0.0007
None ($M = 19.4, SD = 19.0$) > Multiple ($M = 7.43, SD = 9.98$)	< 0.0001
(Other comparisons were not significant)	

Table A.10: ANCOVA results for input area correctness (proportion correct after removing “not sure” responses). The experimental design improved correctness, with participants in the experimental group performing more accurately than those in the traditional group. Academic field also predicted higher correctness, whereas other covariates (e.g., gender, age, disability, statistics course experience) had minimal impact. Uncertainty was measured in the input area, but because only two coding questions were used, values were limited to 0%, 50%, or 100% and were not suitable for ANOVA analysis; see the main text for the descriptive results of input area uncertainty.

Source	Uncertainty					Correctness				
	SS	df	F	p	η_p^2	SS	df	F	p	η_p^2
Design	-	-	-	-	-	5948	1	12.021	< 0.001***	0.070
Gender	-	-	-	-	-	1231	1	2.488	0.117	0.010
Academic Field	-	-	-	-	-	2361	1	4.772	0.030*	0.030
Primary Language	-	-	-	-	-	770	1	1.556	0.214	0.009
Age Group	-	-	-	-	-	1018	1	2.057	0.153	0.010
Has Disability	-	-	-	-	-	356	1	0.720	0.397	0.004
Statistics Course Experience	-	-	-	-	-	681	2	0.688	0.504	0.008
Anxiety	-	-	-	-	-	1	1	0.003	0.957	0.000
Confidence	-	-	-	-	-	3009	1	6.082	0.015*	0.040
Design $\times$ Gender	-	-	-	-	-	953	1	1.926	0.167	0.010
Gender $\times$ Academic Field	-	-	-	-	-	104	1	0.210	0.648	0.001
Residuals	-	-				81146	164			

A.2.5 Output Area

Table A.11: Input area ANCOVA results for overall and code-related ratings of anxiety and confidence. Design had a positive effect on participants' confidence across both overall and code-related ratings and reduced anxiety specifically for code-related items (e.g., format, naming convention, amount and syntax of the code), indicating that the experimental design helped participants feel more confident and less anxious when entering statistical code. Several effects emerged for covariates: gender and age were associated with overall anxiety, while statistics course experience and academic field were linked to code-related confidence and anxiety. These results suggest that design improvements can selectively enhance participants' experience with code entry tasks, particularly in reducing anxiety related to understanding and using statistical code.

Source	Anxiety					Confidence				
	SS	df	F	p	η_p^2	SS	df	F	p	η_p^2
<i>Overall Ratings</i>										
Design	8.67	1	3.53	0.061	0.010	18.58	1	8.25	0.004**	0.030
Gender	11.92	1	4.86	0.028*	0.020	10.49	1	4.66	0.032*	0.020
Academic Field	3.22	1	1.31	0.253	0.005	11.91	1	5.29	0.022*	0.020
Primary Lang	1.38	1	0.56	0.454	0.002	6.27	1	2.78	0.097	0.010
Age Group	12.60	1	5.14	0.024*	0.020	10.20	1	4.53	0.034*	0.020
Has Disability	1.65	1	0.67	0.413	0.003	0.04	1	0.02	0.892	0.000
Stats Course Exp	12.58	2	2.57	0.079	0.020	17.84	2	3.96	0.020*	0.030
Design $\times$ Gender	3.74	1	1.53	0.218	0.006	0.15	1	0.07	0.798	0.000
Gender $\times$ Field	3.01	1	1.23	0.269	0.005	14.42	1	6.40	0.012*	0.020
Residuals	637.52	260				585.53	260			
<i>Code-Related Ratings</i>										
Design	19.27	1	8.12	0.005**	0.030	17.64	1	7.49	0.007**	0.030
Gender	6.89	1	2.90	0.090	0.010	14.31	1	6.07	0.014*	0.020
Academic Field	9.65	1	4.07	0.045*	0.020	22.65	1	9.61	0.002**	0.040
Primary Lang	2.73	1	1.15	0.284	0.004	5.85	1	2.48	0.116	0.009
Age Group	12.52	1	5.27	0.022*	0.020	8.81	1	3.74	0.054	0.010
Has Disability	1.19	1	0.50	0.480	0.002	0.73	1	0.31	0.579	0.001
Stats Course Exp	10.98	2	2.31	0.101	0.020	38.43	2	8.15	< 0.001***	0.060
Design $\times$ Gender	11.33	1	4.77	0.030 *	0.020	0.82	1	0.35	0.555	0.001
Gender $\times$ Field	5.92	1	2.49	0.116	0.010	21.17	1	8.98	0.003**	0.030
Residuals	617.13	260				612.71	260			

Table A.12: Tukey–Kramer post hoc comparisons for input area. Having taken at least one statistics course consistently helped promote higher confidence in the input area. Women outside of formal sciences felt significantly less confident than all other participants.

Comparison	p-value
<i>Design × Gender (Anxiety: Code-Related Ratings)</i>	
Traditional Male ($M = 3.30, SD = 1.72$) > Experimental Male ($M = 2.49, SD = 1.37$)	0.0023
(Other comparisons were not significant)	
<i>Statistics Course Experience (Confidence: Overall Ratings)</i>	
None ($M = 2.57, SD = 1.57$) < Multiple ($M = 3.30, SD = 1.79$)	0.0151
(Other comparisons were not significant)	
<i>Gender × Academic Field (Confidence: Overall Ratings)</i>	
Female Formal Sciences ($M = 3.30, SD = 1.77$) > Female Other ($M = 2.19, SD = 1.38$)	0.0046
Male Formal Sciences ($M = 3.04, SD = 1.52$) > Female Other ($M = 2.19, SD = 1.38$)	0.0002
Male Other ($M = 3.06, SD = 1.66$) > Female Other ($M = 2.19, SD = 1.38$)	0.0064
(Other comparisons were not significant)	

Table A.13: Task 3 correctness when attempted for each criterion. See the Psychometrics section for a full description.

Task 3 Criteria (short description)	Experimental		Traditional	
	Criteria #	% Correct	Criteria #	% Correct
Data frame object created	1	93%	1	95%
File name correct	2	92%	2	89%
File includes quotations	3	95%	3	89%
File includes .csv	4	97%	4	97%
Input factor set correctly	5	75%	5	47%
Output variable set correctly	6	90%	6	61%
Main test & parameters	7	84%	7	40%
Assumption/posthoc/effect size	8	92%	8	76%
Syntax correct	9	68%	9	42%
No extra tests/factors/variables	10	83%	10	60%
No missing tests/factors/variables	11	85%	11	55%
Code produces correct results	12	57%	12	29%

Table A.14: Task 4 correctness when attempted for each criterion. See the Psychometrics section for a full description.

Task 4 Criteria (short description)	Experimental		Traditional	
	Criteria #	% Correct	Criteria #	% Correct
Data frame object created	1	93%	1	97%
File name correct	2	91%	2	93%
File includes quotations	3	97%	3	97%
File includes .csv	4	96%	4	99%
Input factor formatted	5	88%	-	-
Output variable set correctly	-	-	5	84%
Repeated factor set	6	75%	6	97%
Repeated factor list named	7	84%	7	88%
Repeated factor list formatted	8	74%	8	96%
Subject identifier set	-	-	9	47%
Input variable set correctly	9	94%	10	87%
Repeated variable set correctly	10	94%	11	74%
Main test & parameters	11	87%	12	43%
Assumption/posthoc/effect size	12	96%	13	93%
Syntax correct	13	60%	14	43%
No extra tests/factors/variables	14	91%	15	62%
No missing tests/factors/variables	15	83%	16	71%
Code produces correct results	16	49%	17	41%

Table A.15: Output area ANCOVA results for overall ratings of anxiety and confidence. Design significantly affected anxiety and confidence, with experimental participants showing lower anxiety and higher confidence. Covariate effects were mostly nonsignificant.

Source	Anxiety					Confidence				
	SS	df	F	p	η_p^2	SS	df	F	p	η_p^2
Design	14.67	1	6.362	0.012*	0.020	25.14	1	12.810	< 0.001***	0.050
Gender	0.85	1	0.370	0.544	0.001	3.52	1	1.795	0.181	0.007
Academic Field	0.02	1	0.010	0.919	0.000	0.03	1	0.016	0.898	0.000
Primary Lang	0.21	1	0.091	0.763	0.000	2.54	1	1.297	0.256	0.005
Age Group	3.60	1	1.562	0.213	0.006	3.67	1	1.870	0.173	0.007
Has Disability	6.49	1	2.816	0.095	0.010	2.53	1	1.291	0.257	0.005
Stats Course	3.16	2	0.685	0.505	0.005	21.39	2	5.451	0.005**	0.040
Exp										
Design $\times$ Gender	1.16	1	0.504	0.478	0.002	2.73	1	1.391	0.239	0.005
Gender $\times$ Field	1.40	1	0.607	0.437	0.002	6.70	1	3.414	0.066	0.010
Residuals	592.72	257				504.35	257			

Table A.16: ANCOVA results for output area uncertainty (proportion of “not sure” responses across all questions) and correctness (calculated after excluding “not sure” responses). The experimental design reduced uncertainty and improved accuracy. Confidence strongly predicted both, while course experience reduced uncertainty but was less clearly tied to accuracy.

Source	Uncertainty					Correctness				
	SS	df	F	p	η_p^2	SS	df	F	p	η_p^2
Design	23545	1	32.80	< .001***	0.110	3933	1	7.144	0.008**	0.030
Gender	616	1	0.858	0.355	0.003	234	1	0.424	0.516	0.002
Academic Field	192	1	0.268	0.605	0.001	263	1	0.478	0.490	0.002
Primary Lang	501	1	0.698	0.404	0.003	159	1	0.289	0.591	0.001
Age Group	2064	1	2.875	0.091	0.010	6653	1	12.09	< .001***	0.050
Has Disability	1404	1	1.957	0.163	0.008	218	1	0.397	0.529	0.002
Stats Course	10746	2	7.486	< .001***	0.060	2619	2	2.378	0.095	0.020
Exp										
Anxiety	520	1	0.725	0.395	0.003	55	1	0.099	0.753	0.000
Confidence	51291	1	71.46	< .001***	0.220	3575	1	6.494	0.012*	0.030
Design $\times$ Gender	40	1	0.056	0.814	0.000	239	1	0.434	0.511	0.002
Gender $\times$ Field	612	1	0.853	0.357	0.003	8	1	0.015	0.903	0.000
Residuals	181595	253				118911	216			

Appendix B

Study 2

B.1 Tasks

Task 1: Pearson Correlation

Task Description. “A researcher is interested in the relationship between sleep and concentration in college students. He measures how many hours each student slept and their reported concentration levels, and wants to determine whether these two measures are related. Use the dataset he collected to set up the appropriate analysis to determine if the two measures are related.”

Dataset. The dataset contains 30 observations across three variables: a numeric subject identifier (*student_id*), a continuous numeric predictor (*hours_of_sleep*), and a continuous numeric outcome (*level_of_concentration*). A representative excerpt is shown in Table B.1.

Table B.1: Task 1 dataset excerpt (first 8 rows of 30).

student_id	hours_of_sleep	level_of_concentration
1	4.3	62
2	6.2	70
3	5.1	65
4	7.4	78
5	8.1	82
6	6.7	74
7	5.6	68
8	7.2	76

Rationale. Task 1 introduces the most fundamental statistical relationship, a bivariate correlation between two continuous numeric variables. It was selected as the opening task because correlation requires the fewest conceptual decisions of any task in the sequence: two numeric variables, no grouping factor, no repeated measures, and no interaction term, minimizing cognitive load at first encounter with an unfamiliar interface. The *student_id* variable is included as an optional identifier, providing a realistic dataset structure without being required for the analysis. For the purposes of this study, *hours_of_sleep* was designated as the input variable and *level_of_concentration* as the output variable, consistent with the directional framing of the task description. While correlation is mathematically symmetric, a consistent role assignment was required for grading purposes, and both interfaces accepted only the designated ordering.

Control Interface Answer

```
pg.corr(x=df["hours_of_sleep"], y=df["level_of_concentration"])
```

No explicit type declarations are required as both variables default to numeric in the dataset. The *student_id* variable should not be included.

Experimental Interface Answer

```
design:AddVariable(frame>CreateNumericalInputVariable("hours_of_sleep"))
design:AddVariable(frame>CreateNumericalOutputVariable("level_of_concentration"))
StatisticalTest test = frame:GetTest(design)
TestResult result = test:Run()
```

The *student_id* variable may optionally be included as a *CreateSubjectIdentifierVariable* but is not required for a valid solution.

Task 2: One-Way Analysis of Variance

Task Description. “A teacher wants to evaluate differences in exam performance across study methods. She separates her students into three groups: one uses flash cards, one completes a practice exam, and one receives no study material. Use the dataset she collected to set up the appropriate analysis to compare exam scores across the groups.”

Dataset. The dataset contains 30 observations across three variables: a numeric subject identifier (*student_id*), a categorical between-subjects factor (*study_method*), and a continuous numeric outcome (*exam_score*). A representative excerpt is shown in Table B.2.

Rationale. Task 2 introduces the one-way between-subjects ANOVA, extending the two-variable structure of Task 1 by using a categorical grouping factor. The key new concept introduced here is the input role

Table B.2: Task 2 dataset excerpt (first 9 rows of 30).

<code>student_id</code>	<code>study_method</code>	<code>exam_score</code>
1	Flash cards	81
2	Practice exam	73
3	None	66
4	Flash cards	88
5	Practice exam	78
6	None	71
7	Flash cards	79
8	Practice exam	76
9	None	69

for a categorical variable, *study_method* is a between-subjects factor whose levels define the groups being compared, while *exam_score* remains the numeric outcome. This task also introduces the first opportunity for the type distinction to matter: *study_method* must be recognized as categorical rather than numeric.

Control Interface Answer

```
pg.anova(data=df, dv="exam_score", between="study_method")
```

The *pg.anova()* block enforces categorical type for the *between* parameter and default (numeric) type for the *dv* parameter, meaning explicit *astype()* declarations are not strictly required, the system overrides variable types based on parameter placement. The *student_id* variable should not be included.

Experimental Interface Answer

```
design:AddVariable(frame>CreateCategoricalInputVariable("study_method"))
design:AddVariable(frame>CreateNumericalOutputVariable("exam_score"))
StatisticalTest test = frame:GetTest(design)
TestResult result = test:Run()
```

The *student_id* variable may optionally be included as a *CreateSubjectIdentifierVariable* but is not required for a valid solution.

Task 3: Chi-Squared Test of Independence

Task Description. “A school counselor wants to investigate whether students’ preferred learning style (visual, auditory, kinesthetic) is related to whether they attend tutoring sessions. She collects data on students and wants to test whether these two characteristics are independent.”

Dataset. The dataset contains 30 observations across three variables: a numeric subject identifier (*student_id*), a nominal categorical variable (*learning_style*), and a binary categorical outcome (*tutoring_attendance*). A representative excerpt is shown in Table B.3.

Table B.3: Task 3 dataset excerpt (first 9 rows of 30).

<i>student_id</i>	<i>learning_style</i>	<i>tutoring_attendance</i>
1	Visual	Yes
2	Auditory	No
3	Kinesthetic	Yes
4	Visual	Yes
5	Auditory	Yes
6	Kinesthetic	No
7	Visual	No
8	Auditory	No
9	Kinesthetic	Yes

Rationale. Task 3 introduces the chi-squared test of independence, the first task in the sequence involving two categorical variables with no continuous numeric outcome. The key new concept is that both variables are categorical, *tutoring_attendance* is designated as the output variable by convention, consistent with the task framing of attendance as the characteristic being predicted by learning style. As with Task 2, the *student_id* variable is included as an optional identifier not required for the analysis.

Control Interface Answer

```
pg.chi2_independence(data=df, x=df["learning_style"], y=df["tutoring_attendance"])
```

The `pg.chi2_independence()` block enforces categorical type on both the *x* and *y* parameters, meaning explicit `astype()` declarations are not required. The *student_id* variable should not be included.

Experimental Interface Answer

```
design:AddVariable(frame>CreateCategoricalInputVariable("learning_style"))
design:AddVariable(frame>CreateCategoricalOutputVariable("tutoring_attendance"))
StatisticalTest test = frame:GetTest(design)
TestResult result = test:Run()
```

The *student_id* variable may optionally be included as a *CreateSubjectIdentifierVariable* but is not required for a valid solution.

Task 4: Two-Way Analysis of Variance

Task Description. “A researcher examines whether study method (flash cards vs. practice exam) and study environment (quiet room vs. noisy room) together influence test performance. Students are assigned to different combinations, and the researcher wants to compare performance across all groups.”

Dataset. The dataset contains 20 observations across four variables: a numeric subject identifier (*student_id*), two categorical between-subjects factors (*study_method* and *study_environment*), and a continuous numeric outcome (*test_score*). Each cell of the 2×2 factorial design contains five observations. A representative excerpt is shown in Table B.4.

Table B.4: Task 4 dataset excerpt (first 8 rows of 20).

student_id	study_method	study_environment	test_score
1	Flash cards	Quiet room	88
2	Flash cards	Quiet room	85
3	Flash cards	Quiet room	90
4	Flash cards	Quiet room	87
5	Flash cards	Quiet room	89
6	Flash cards	Noisy room	78
7	Flash cards	Noisy room	75
8	Flash cards	Noisy room	80

Rationale. Task 4 introduces two new concepts simultaneously: a second between-subjects input factor and an interaction term between the two factors. The task is the first in the sequence to require an explicit interaction specification, testing whether participants recognize that a factorial design involves not only the main effects of each factor but also their combination. Both *study_method* and *study_environment* are categorical input variables, and *test_score* is the continuous numeric output. The *student_id* variable is included as an optional identifier. The task description uses the phrase “together influence,” which is intended to cue the interactive nature of the design without explicitly naming an interaction term.

Control Interface Answer

```
pg.anova(data=df, dv="test_score", between=["study_method", "study_environment"])
```

The two-way ANOVA block (*pg.anova* with a list *between*=[] parameter) distinguishes itself from the one-way ANOVA block by accepting two variables in the *between* parameter. The *pg.anova()* block enforces categorical type on both *between* variables and default type on the *dv*, meaning explicit *astype()* declarations are not required. The interaction between the two factors is implicitly specified by including both variables

in the *between* list and does not require a separate interaction term. The *student_id* variable should not be included.

Experimental Interface Answer

```
design:AddInputInteraction(  
    frame>CreateCategoricalInputVariable("study_method"),  
    frame>CreateCategoricalInputVariable("study_environment"))  
design:AddVariable(frame>CreateNumericalOutputVariable("test_score"))  
StatisticalTest test = frame:GetTest(design)  
TestResult result = test:Run()
```

The interaction block (*AddInputInteraction*) implicitly carries both input variables, making standalone *AddVariable* blocks for *study_method* and *study_environment* optional. If included, they must be classified correctly as *CreateCategoricalInputVariable*. The *student_id* variable may optionally be included as a *CreateSubjectIdentifierVariable* but is not required for a valid solution.

Task 5: One-Way Repeated Measures ANOVA

Task Description. “A cognitive psychologist measures participants’ reaction times under three different lighting conditions: bright, dim, and normal. Each participant repeats tasks under all three conditions, and the psychologist wants to determine whether reaction times differ under these lighting conditions.”

Dataset. The dataset contains 30 observations across three variables: a categorical subject identifier (*participant_id*), a categorical within-subjects factor (*lighting_condition*), and a continuous numeric outcome (*reaction_time*). Each of the 10 participants contributes three observations, one per lighting condition. A representative excerpt is shown in Table B.5.

Rationale. Task 5 introduces the repeated measures design, the most conceptually novel element of the task sequence thus far. The within-subjects factor *lighting_condition* is the first variable in the sequence to require the repeat role, distinct from the input role used for between-subjects factors in Tasks 2 and 4. Recognizing the repeat role requires understanding that the same participant appears multiple times in the dataset under different conditions, which is visually apparent from the dataset structure (participant 1 appears in rows 1–3, participant 2 in rows 4–6, etc.). Unlike Tasks 1–4, *participant_id* is a required variable in this task, serving as the subject identifier necessary to link repeated observations to individual participants.

Control Interface Answer

Table B.5: Task 5 dataset excerpt (first 9 rows of 30).

participant_id	lighting_condition	reaction_time
1	Bright	320
1	Dim	340
1	Normal	310
2	Bright	305
2	Dim	325
2	Normal	300
3	Bright	330
3	Dim	345
3	Normal	315

```
pg.rm_anova(data=df, dv="reaction_time", within="lighting_condition",
            subject="participant_id")
```

The `pg.rm_anova()` block enforces categorical type on the *within* and *subject* parameters and default type on the *dv* parameter, meaning explicit *astype()* declarations are not required. All three variables are required and must be assigned to their respective parameter slots: *dv*= for the outcome, *within*= for the repeated factor, and *subject*= for the participant identifier.

Experimental Interface Answer

```
design:AddVariable(frame>CreateSubjectIdentifierVariable("participant_id"))
design:AddVariable(frame>CreateCategoricalRepeatedVariable("lighting_condition"))
design:AddVariable(frame>CreateNumericalOutputVariable("reaction_time"))
StatisticalTest test = frame:GetTest(design)
TestResult result = test:Run()
```

All three variables are required. The *lighting_condition* variable may be classified as either *CreateCategoricalRepeatedVariable* or *CreateOrderedRepeatedVariable*. Although the lighting conditions do not have a natural ordering relevant to the research question, ordered categorical treatment is statistically defensible for repeated within-subjects factors, and both classifications were accepted as correct in the experimental interface. The control interface makes no such distinction, `pg.rm_anova()` treats all *within*= variables as categorical regardless of any explicit type declaration.

Task 6: Kruskal-Wallis Test

Task Description. “A nutritionist wants to evaluate client satisfaction across three diet programs. Clients rate their experiences as low (1), medium (2), or high (3), and she wants to determine whether satisfaction

differs across the programs.”

Dataset. The dataset contains 30 observations across three variables: a numeric subject identifier (*client_id*), a categorical between-subjects factor (*diet_program*), and an ordinal outcome (*satisfaction*). Satisfaction is recorded as integer values 1, 2, and 3 representing low, medium, and high ratings respectively. A representative excerpt is shown in Table B.6.

Table B.6: Task 6 dataset excerpt (first 9 rows of 30).

client_id	diet_program	satisfaction
1	Program A	3
2	Program B	2
3	Program C	1
4	Program A	2
5	Program B	3
6	Program C	2
7	Program A	1
8	Program B	2
9	Program C	3

Rationale. Task 6 introduces the Kruskal-Wallis test, the non-parametric analogue of the one-way ANOVA used when the outcome variable is ordinal rather than continuous. The key new concept is the *ordered* type for the *satisfaction* variable, which is stored as integers (1, 2, 3) in the dataset but represents a ranked categorical scale rather than a true continuous measure. Recognizing that satisfaction is ordinal rather than numeric is the primary type challenge of this task. The structure of the task is otherwise similar to Task 2, one categorical input factor and one outcome variable, making the main distinction the type of the outcome. The *client_id* variable is included as an optional identifier.

Control Interface Answer

```
pg.kruskal(data=df, dv="satisfaction", between="diet_program")
```

The `pg.kruskal()` block enforces categorical type on the `between=` slot. The `dv=` slot accepts either the default numeric type or an explicit ordered categorical declaration, as both are statistically valid for a Kruskal-Wallis test on an ordinal outcome. An explicit `astype()` declaration for *satisfaction* is therefore optional. The *client_id* variable should not be included.

Experimental Interface Answer

```

design:AddVariable(frame>CreateCategoricalInputVariable("diet_program"))
design:AddVariable(frame>CreateOrderedOutputVariable("satisfaction"))
StatisticalTest test = frame:GetTest(design)
TestResult result = test:Run()

```

The *client_id* variable may optionally be included as a *CreateSubjectIdentifierVariable* but is not required for a valid solution.

Task 7: Mixed-Design ANOVA

Task Description. “A psychologist studies stress levels in two different job types (office vs. field workers) measuring the same workers at four points during the day: morning, midday, afternoon, and evening. He wants to determine whether stress patterns differ between job types over time.”

Dataset. The dataset contains 40 observations across four variables: a categorical subject identifier (*participant_id*), a categorical between-subjects factor (*job_type*), a categorical within-subjects factor (*time_of_day*), and a continuous numeric outcome (*stress_level*). Each of the 10 participants contributes four observations, one per time point, and participants are nested within job type. A representative excerpt is shown in Table B.7.

Table B.7: Task 7 dataset excerpt (first 8 rows of 40).

participant_id	job_type	time_of_day	stress_level
1	Office	Morning	45
1	Office	Midday	50
1	Office	Afternoon	55
1	Office	Evening	48
2	Office	Morning	42
2	Office	Midday	47
2	Office	Afternoon	52
2	Office	Evening	45

Rationale. Task 7 introduces the mixed-design ANOVA, combining the between-subjects factor from Task 2 (*job_type* as input) with the within-subjects repeated measures factor from Task 5 (*time_of_day* as repeat) and their interaction. It is the first task to require four different variable roles simultaneously: identifier, input, repeat, and output. The interaction between *job_type* and *time_of_day* captures whether stress patterns over the course of the day differ between job types, framed in the task description as whether stress patterns “*differ between job types over time.*”

Control Interface Answer

```
pg.mixed_anova(data=df, dv="stress_level", within="time_of_day",
               between="job_type", subject="participant_id")
```

The `pg.mixed_anova()` block enforces categorical type on the `within=`, `between=`, and `subject=` slots and default type on the `dv=` slot, meaning no explicit `astype()` declarations are required. All four variables are required and must be assigned to their respective parameter slots.

Experimental Interface Answer

```
design:AddVariable(frame>CreateSubjectIdentifierVariable("participant_id"))
design:AddInputInteraction(
    frame>CreateCategoricalInputVariable("job_type"),
    frame>CreateCategoricalRepeatedVariable("time_of_day"))
design:AddVariable(frame>CreateNumericalOutputVariable("stress_level"))
StatisticalTest test = frame:GetTest(design)
TestResult result = test:Run()
```

As in Task 4, the standalone *AddVariable* blocks for *job_type* and *time_of_day* are optional when the interaction block is present, as the interaction implicitly carries both variables. The *participant_id* block is required. The *time_of_day* variable may be classified as either *Categorical* or *Ordered*, consistent with the statistical defensibility of treating ordered time points as ranked within-subjects conditions.

Task 8: Multivariate Analysis of Variance

Task Description. “A school district wants to understand how teaching method (lecture vs. discussion) and class size (small vs. large) influence student outcomes on math scores and reading scores. The district is interested in whether each factor affects performance on its own, as well as whether their combination leads to different outcomes across the scores.”

Dataset. The dataset contains 40 observations across five variables: a numeric subject identifier (*student_id*), two categorical between-subjects factors (*teaching_method* and *class_size*), and two continuous numeric outcomes (*math_score* and *reading_score*). The design is fully crossed with 10 observations per cell. A representative excerpt is shown in Table B.8.

Rationale. Task 8 introduces MANOVA, the first task in the sequence with multiple outcome variables. The key new concept is that both *math_score* and *reading_score* are designated as output variables, requiring participants to recognize that a single analysis can simultaneously model more than one outcome. The two input factors (*teaching_method* and *class_size*) and their interaction follow the same structure as Task 4,

Table B.8: Task 8 dataset excerpt (first 8 rows of 40).

<code>student_id</code>	<code>teaching_method</code>	<code>class_size</code>	<code>math_score</code>	<code>reading_score</code>
1	Lecture	Small	78	74
2	Lecture	Small	80	75
3	Lecture	Small	82	77
4	Lecture	Small	79	76
5	Lecture	Small	81	78
6	Lecture	Large	72	69
7	Lecture	Large	74	71
8	Lecture	Large	70	68

making the primary novel element the presence of two outputs rather than one. This task also marks the first use of the formula syntax in the control group, replacing the named-parameter penguin blocks used in Tasks 1–7.

Control Interface Answer

```
sm.manova(data=df, formula="math_score + reading_score
    ~ teaching_method * class_size")
```

Task 8 is the first task in the control group to use the statsmodels/R formula syntax, in which the `~` operator separates outcomes (left-hand side) from predictors (right-hand side), the `+` operator combines multiple outcomes or predictors, and the `*` operator specifies both main effects and their interaction. Unlike the penguin blocks in Tasks 1–7, the statsmodels blocks do not enforce type overrides through parameter slot assignment; variable types are determined by the dataframe or by explicit type declarations at the time of execution. However, all variables in Task 8 default to their correct types, *teaching_method* and *class_size* as categorical strings, *math_score* and *reading_score* as numeric, meaning no explicit *astype()* declarations are required. The *student_id* variable should not be included.

Experimental Interface Answer

```
design:AddInputInteraction(
    frame>CreateCategoricalInputVariable("teaching_method"),
    frame>CreateCategoricalInputVariable("class_size"))
design:AddVariable(frame>CreateNumericalOutputVariable("math_score"))
design:AddVariable(frame>CreateNumericalOutputVariable("reading_score"))
StatisticalTest test = frame:GetTest(design)
TestResult result = test:Run()
```

As in Tasks 4 and 7, standalone *AddVariable* blocks for *teaching_method* and *class_size* are optional when the interaction block is present. Both output variables (*math_score* and *reading_score*) are required and must be declared individually. The *student_id* variable may optionally be included as a *CreateSubjectIdentifierVariable*.

Task 9: Two-Way Analysis of Covariance

Task Description. “A researcher investigates how exercise program type (cardio vs. strength) and supplement use (yes vs. no) relate to weight loss, taking into account participants’ gender. He wants to know whether differences in weight loss are associated with program type, supplement use, or the combination of the two after considering gender.”

Dataset. The dataset contains 30 observations across five variables: a numeric subject identifier (*participant_id*), two categorical between-subjects factors (*exercise_program* and *supplement*), a categorical covariate (*gender*), and a continuous numeric outcome (*weight_loss*). A representative excerpt is shown in Table B.9.

Table B.9: Task 9 dataset excerpt (first 8 rows of 30).

participant_id	exercise_program	supplement	gender	weight_loss
1	Cardio	Yes	Male	3.2
2	Cardio	Yes	Female	3.8
3	Cardio	Yes	Female	4.0
4	Cardio	Yes	Male	3.5
5	Cardio	Yes	Female	4.1
6	Cardio	No	Male	2.9
7	Cardio	No	Female	3.0
8	Cardio	No	Female	3.2

Rationale. Task 9 introduces the covariate role, the final variable role in the experimental interface’s classification system. The *gender* variable is designated as a control variable, a factor whose influence on the outcome is statistically accounted for rather than directly investigated. The task description signals this through the phrases “taking into account” and “after considering,” which were intended to convey the covariate’s role without using technical terminology. The two input factors (*exercise_program* and *supplement*) and their interaction follow the same structure as Tasks 4 and 8. Although pingouin provides a *pg.ancova()* function, it accepts only numeric covariates and a single between-subjects factor, making it unsuitable for this task which requires a categorical covariate and a two-factor factorial design with

interaction. The statsmodels formula interface was therefore used for the control group.

Control Interface Answer

```
sm.anova_lm.ols(data=df, formula="weight_loss
    ~ exercise_program * supplement + gender")
```

In the statsmodels formula syntax, the covariate *gender* is specified using the $+$ operator on the right-hand side of the formula, following the interaction term. Variables default to their correct types in the dataset, meaning no explicit *astype()* declarations are required. The *participant_id* variable should not be included.

Experimental Interface Answer

```
design:AddInputInteraction(
    frame>CreateCategoricalInputVariable("exercise_program"),
    frame>CreateCategoricalInputVariable("supplement"))
design:AddVariable(frame>CreateCategoricalControlVariable("gender"))
design:AddVariable(frame>CreateNumericalOutputVariable("weight_loss"))
StatisticalTest test = frame:GetTest(design)
TestResult result = test:Run()
```

As in Tasks 4, 7, and 8, standalone *AddVariable* blocks for *exercise_program* and *supplement* are optional when the interaction block is present. The *participant_id* variable may optionally be included as a *CreateSubjectIdentifierVariable*.

Task 10: Mixed Analysis of Covariance

Task Description. “A rehabilitation specialist compares two therapy types. Patient mobility is measured at three distinct treatment phases: pre-treatment, mid-treatment, and post-treatment. She wants to compare mobility across these phases and also examine whether the effect of therapy type differs by phase, all while controlling for age.”

Dataset. The dataset contains 90 observations across five variables: a numeric subject identifier (*patient_id*), a between-subjects factor (*therapy_type*), a continuous numeric covariate (*age*), a within-subjects repeated factor (*treatment_phase*), and a continuous numeric outcome (*mobility*). Each of the 30 patients contributes three observations, one per treatment phase. A representative excerpt is shown in Table B.10.

Rationale. Task 10 is the most conceptually complex task in the sequence, combining the between-subjects input role (*therapy_type*), the repeated measures role (*treatment_phase*), the covariate role (*age*), and their interaction, requiring all five variable roles simultaneously for the first time. Two deliberate design decisions

Table B.10: Task 10 dataset excerpt (first 6 rows of 90).

patient_id	therapy_type	age	treatment_phase	mobility
1	1	65	Pre-treatment	50
1	1	65	Mid-treatment	55
1	1	65	Post-treatment	60
2	1	72	Pre-treatment	48
2	1	72	Mid-treatment	53
2	1	72	Post-treatment	58

are notable. First, *therapy_type* is stored as integer values (1 and 2) rather than string labels, unlike all other categorical variables in the task sequence which use string representations. This was intentional, correctly identifying *therapy_type* as categorical despite its numeric appearance requires genuine understanding of the distinction between a variable's storage format and its statistical type. Failing to recognize this distinction has meaningful analytic consequences: treating an integer-coded categorical variable as continuous causes the model to interpret the arbitrary numeric codes (1, 2) as a meaningful scale, producing a linear trend estimate rather than a group comparison. This type of error is common in practice and represents a genuine threat to the validity of statistical conclusions. Second, *age* is a continuous numeric covariate, in contrast to the categorical covariate (*gender*) introduced in Task 9, demonstrating that the control role is not type-restricted.

Control Interface Answer

```
df["therapy_type"] = df["therapy_type"].astype("category", ordered=False)
sm.mixedlm(data=df, formula="mobility ~ therapy_type * treatment_phase + age",
           groups=df["patient_id"])
```

Unlike all other variables in the task sequence, *therapy_type* defaults to numeric in the dataset due to its integer representation. An explicit *astype()* declaration to categorical is required for a correct solution, as leaving *therapy_type* as numeric would cause it to be treated as a continuous predictor rather than a categorical factor in the mixed effects model. The *groups=* parameter specifies the subject identifier for the random effects structure and accepts numeric or categorical type. The *treatment_phase* variable may be classified as either *ordered=False* or *ordered=True* but no explicit *astype()* declaration is required.

Experimental Interface Answer

```
design:AddVariable(frame>CreateSubjectIdentifierVariable("patient_id"))
design:AddInputInteraction(
    frame>CreateCategoricalInputVariable("therapy_type"),
    frame>CreateCategoricalRepeatedVariable("treatment_phase"))
```

```

design:AddVariable(frame>CreateNumericalControlVariable("age"))
design:AddVariable(frame>CreateNumericalOutputVariable("mobility"))
StatisticalTest test = frame:GetTest(design)
TestResult result = test:Run()

```

As in previous interaction tasks, standalone *AddVariable* blocks for *therapy_type* and *treatment_phase* are optional when the interaction block is present. The *treatment_phase* variable may be classified as either *CreateCategoricalRepeatedVariable* or *CreateOrderedRepeatedVariable*.

Task 11: Linear Regression

Task Description. “A city planner wants to predict house prices based on several characteristics, including square footage, lot size, distance to downtown, and the neighborhood in which the house is located. He wants to build a model that can predict price from the collected data.”

Dataset. The dataset contains 40 observations across six variables: a numeric subject identifier (*house_id*), three continuous numeric predictors (*square_footage*, *lot_size_sqft*, *distance_to_downtown*), one nominal categorical predictor (*neighborhood*), and a continuous numeric outcome (*price*). A representative excerpt is shown in Table B.11.

Table B.11: Task 11 dataset excerpt (first 6 rows of 40).

house_id	square_footage	lot_size_sqft	neighborhood	dist_to_downtown	price
1	2100	5420.5	North	5.2	460000
2	1800	4850.2	South	8.1	400000
3	2200	6100.8	East	3.5	500000
4	1600	4200.3	West	10.3	380000
5	2500	6800.1	North	4.7	550000
6	1900	5100.7	South	7.4	420000

Rationale. Task 11 introduces linear regression, marking a paradigm shift from statistical hypothesis testing to predictive modeling. In the experimental interface this is reflected by replacing the *GetTest()* block with a *GetModel()* block, while in the control interface it introduces the sklearn preprocessing and model-fitting syntax in place of pingouin and statsmodels. The task description uses the word *predict* and *model*, which were intended to signal the shift to a modeling paradigm. The task introduces preprocessing requirements for the first time for the control group. The three continuous numeric predictors require normalization (*StandardScaler*) due to their differing scales, *square_footage* ranges in the thousands while *distance_to_downtown* ranges in the single digits. The categorical predictor *neighborhood* requires one-hot

encoding (*OneHotEncoder*) to be represented numerically for the regression model. The outcome variable *price* does not require preprocessing.

Control Interface Answer

```
df["square_footage"] = StandardScaler(df["square_footage"])
df["lot_size_sqft"] = StandardScaler(df["lot_size_sqft"])
df["distance_to_downtown"] = StandardScaler(df["distance_to_downtown"])
df["neighborhood"] = OneHotEncoder(df["neighborhood"])
X = df[square_footage, lot_size_sqft, neighborhood, distance_to_downtown]
y = df[price]
skl.LinearRegression().fit(X, y)
```

The control interface requires participants to explicitly apply preprocessing to each variable before fitting the model. Preprocessing blocks (*StandardScaler*, *OneHotEncoder*, *OrdinalEncoder*, *LabelEncoder*) become available in the palette for Tasks 11 and 12 only, as *pingouin* and *statsmodels* apply these preprocessing steps internally. The *price* outcome variable does not require preprocessing and should not have a preprocessing block applied. The *house_id* variable should not be included.

Experimental Interface Answer

```
design:AddVariable(frame>CreateNumericalInputVariable("square_footage"))
design:AddVariable(frame>CreateNumericalInputVariable("lot_size_sqft"))
design:AddVariable(frame>CreateCategoricalInputVariable("neighborhood"))
design:AddVariable(frame>CreateNumericalInputVariable("distance_to_downtown"))
design:AddVariable(frame>CreateNumericalOutputVariable("price"))
UntrainedModel model = frame:GetModel(design)
TrainedModel trained = model:Train()
```

The *house_id* variable may optionally be included as a *CreateSubjectIdentifierVariable* but is not required. In the experimental interface, preprocessing is handled automatically by the system once variable roles and types are correctly specified; participants are not required to select or apply preprocessing steps explicitly.

Task 12: Logistic Regression

Task Description. “A university has collected orientation assessment data from students over the past five years, along with the major each student successfully completed (*Business*, *Humanities*, *STEM*, or *Health Sciences*). They want to use this historical data to build a predictive model that recommends which major incoming students are likely to succeed in based on their orientation assessment scores.”

Dataset. The dataset contains 36 observations across six variables: a numeric subject identifier (*student_id*), four assessment score predictors (*quantitative_score*, *verbal_score*, *abstract_reasoning*, *interpersonal_score*), and a nominal categorical outcome (*major*). A representative excerpt is shown in Table B.12.

Table B.12: Task 12 dataset excerpt (first 6 rows of 36).

student_id	quant_score	verbal_score	abs_reasoning	interp_score	major
1	60	4	8	5	Business
2	72	5	10	5	Business
3	65	4	12	4	Business
10	65	5	12	4	Humanities
19	92	2	25	2	STEM
28	70	4	10	5	Health Sciences

Rationale. Task 12 introduces logistic regression, the classification analogue of the linear regression introduced in Task 11. The key distinction from Task 11 is the nature of the outcome variable: *major* is a nominal categorical variable with four classes rather than a continuous numeric variable. This requires participants to recognize that predicting a categorical outcome calls for a classification model rather than a regression model, and in the control interface, that the outcome variable requires label encoding (*LabelEncoder*) rather than normalization. The predictor variables *verbal_score* and *interpersonal_score* use a narrow integer scale (2–5) while *quantitative_score* and *abstract_reasoning* use broader ranges, making normalization necessary across all four predictors. The *student_id* variable is included as an optional identifier and does not require preprocessing.

Control Interface Answer

```
df["quantitative_score"] = StandardScaler(df["quantitative_score"])
df["verbal_score"] = StandardScaler(df["verbal_score"])
df["abstract_reasoning"] = StandardScaler(df["abstract_reasoning"])
df["interpersonal_score"] = StandardScaler(df["interpersonal_score"])
X = df[quantitative_score, verbal_score, abstract_reasoning, interpersonal_score]
y = df[major]
skl.LogisticRegression().fit(X, y)
```

All four numeric predictors require normalization via *StandardScaler*. The categorical outcome *major* does not require preprocessing as *LogisticRegression* can handle the encoding of string columns itself. If preprocessing is applied to *major*, it requires *LabelEncoder* rather than *OneHotEncoder*, as label encoding is

appropriate for the target variable in a multiclass classification model. The *student_id* variable should not be included and does not require preprocessing.

Experimental Interface Answer

```
design:AddVariable(frame>CreateNumericalInputVariable("quantitative_score"))
design:AddVariable(frame>CreateNumericalInputVariable("verbal_score"))
design:AddVariable(frame>CreateNumericalInputVariable("abstract_reasoning"))
design:AddVariable(frame>CreateNumericalInputVariable("interpersonal_score"))
design:AddVariable(frame>CreateCategoricalOutputVariable("major"))
UntrainedModel model = frame:GetModel(design)
TrainedModel trained = model:Train()
```

The *student_id* variable may optionally be included as a *CreateSubjectIdentifierVariable* but is not required. As in Task 11, preprocessing is handled automatically by the system once variable roles and types are correctly specified. The system resolves logistic regression as the appropriate model given the categorical output variable, without requiring participants to identify or name the model explicitly.

Bibliography

- [1] Alan Baddeley. “Working Memory”. In: *Science* 255.5044 (1992), pp. 556–559. DOI: 10.1126/science.1736359. eprint: <https://www.science.org/doi/pdf/10.1126/science.1736359>. <https://www.science.org/doi/abs/10.1126/science.1736359>.
- [2] M. Bakker and J. M. Wicherts. “The (mis)reporting of statistical results in psychology journals.” In: *Behavior Research Methods* 43 (2011), pp. 666–678. <https://doi.org/10.3758/s13428-011-0089-5>.
- [3] Mustafa Baloglu. “Individual differences in statistics anxiety among college students”. en. In: *Personality and Individual Differences* 34.5 (Apr. 2003), pp. 855–865. ISSN: 01918869. DOI: 10.1016/S0191-8869(02)00076-4. <https://linkinghub.elsevier.com/retrieve/pii/S0191886902000764> (visited on 04/08/2025).
- [4] Mustafa Baloglu, M. Engin Deniz, and Şahin Kesici. “A descriptive study of individual and cross-cultural differences in statistics anxiety”. In: *Learning and Individual Differences* 21.4 (2011), pp. 387–391. ISSN: 1041-6080. DOI: <https://doi.org/10.1016/j.lindif.2011.03.003>. <https://www.sciencedirect.com/science/article/pii/S1041608011000380>.
- [5] Mustafa Baloglu and Paul Zelhart. “Statistical Anxiety: A Detailed Review of the Literature”. In: *Psychology and Education* 40 (Jan. 2003), pp. 27–37. DOI: 10.1007/BF03341649.
- [6] Albert Bandura. *Self-efficacy: the exercise of control*. 1st ed. New York: W.H. Freeman and Company, 1997.
- [7] Albert Bandura. “Self-efficacy: Toward a unifying theory of behavioral change.” en. In: *Psychological Review* 84.2 (1977), pp. 191–215. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/0033-295X.84.2.191. <https://doi.apa.org/doi/10.1037/0033-295X.84.2.191> (visited on 04/10/2026).
- [8] Laura Beckwith et al. “On to the Real World: Gender and Self-Efficacy in Excel”. In: *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing*. VLHCC ’07. USA: IEEE Computer Society, 2007, pp. 119–126. ISBN: 0769529879. DOI: 10.1109/VLHCC.2007.42. <https://doi.org/10.1109/VLHCC.2007.42>.

- [9] Wicher Bergsma. “A bias-correction for Cramér’s and Tschuprow’s”. en. In: *Journal of the Korean Statistical Society* 42.3 (Sept. 2013), pp. 323–328. ISSN: 12263192. DOI: 10.1016/j.jkss.2012.10.002. <https://linkinghub.elsevier.com/retrieve/pii/S1226319212001032> (visited on 04/04/2024).
- [10] Kenneth J. Berry, Janis E. Johnston, and Paul W. Mielke. “An Alternative Measure of Effect Size for Cochran’s Q Test for Related Proportions”. en. In: *Perceptual and Motor Skills* 104.3_suppl (June 2007), pp. 1236–1242. ISSN: 0031-5125, 1558-688X. DOI: 10.2466/pms.104.4.1236-1242. <http://journals.sagepub.com/doi/10.2466/pms.104.4.1236-1242> (visited on 04/04/2024).
- [11] Sabine Beurze et al. “Statistics Anxiety: A Barrier for Education in Research Methodology for Medical Students?” In: *Medical Science Educator* 23 (Sept. 2014), pp. 377–384. DOI: 10.1007/BF03341649.
- [12] Menucha Birenbaum and Shoshana Eylath. “Who is afraid of statistics? Correlates of statistics anxiety among students of educational sciences”. en. In: *Educational Research* 36.1 (Mar. 1994), pp. 93–98. ISSN: 0013-1881, 1469-5847. DOI: 10.1080/0013188940360110. <https://www.tandfonline.com/doi/full/10.1080/0013188940360110> (visited on 04/02/2025).
- [13] Carlo E. Bonferroni. “Teoria statistica delle classi e calcolo delle probabilità”. In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8 (1936), pp. 3–62.
- [14] Albert H. Bowker. “A Test for Symmetry in Contingency Tables”. en. In: *Journal of the American Statistical Association* 43.244 (Dec. 1948), pp. 572–574. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.1948.10483284. <http://www.tandfonline.com/doi/abs/10.1080/01621459.1948.10483284> (visited on 04/04/2024).
- [15] Virginia Braun and Victoria Clarke. “Using thematic analysis in psychology”. In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa. eprint: <https://doi.org/10.1191/1478088706qp063oa>. <https://doi.org/10.1191/1478088706qp063oa>.
- [16] Bureau of Labor Statistics. *Occupational Outlook Handbook: Mathematicians and Statisticians*. Accessed: 2024-06-01. 2023. <https://www.bls.gov/ooh/math/mathematicians-and-statisticians.htm>.
- [17] Daniel Casasanto. “Embodiment of abstract concepts: Good and bad in right- and left-handers”. In: *Journal of Experimental Psychology: General* 138.3 (2009), pp. 351–367. DOI: 10.1037/a0015854.
- [18] P. K. Chew and D. B. Dillon. “Statistics Anxiety Update: Refining the Construct and Recommendations for a New Research Agenda.” In: *Perspectives on psychological science : a journal of the Association for Psychological Science* 9.2 (2014), pp. 196–208. <https://doi.org/10.1177/1745691613518077>.
- [19] Francesca Chiesi and Caterina Primi. “Assessing statistics attitudes among college students: Psychometric properties of the Italian version of the Survey of Attitudes toward Statistics (SATS)”. en. In: *Learning and Individual Differences* 19.2 (June 2009), pp. 309–313. ISSN: 10416080. DOI: 10.1016/j.lindif.2008.10.008. <https://linkinghub.elsevier.com/retrieve/pii/S1041608008001106> (visited on 02/12/2025).

- [20] S. K. Chowdhury. “Prior Knowledge, Sex and Students’ Attitude towards Statistics: A Study on Postgraduate Education Students.” In: *American Journal of Educational Research* 6.3 (2018), pp. 270–276. DOI: 10.12691/education-6-3-14. <http://pubs.sciepub.com/education/6/3/14>.
- [21] W. G. Cochran. “THE COMPARISON OF PERCENTAGES IN MATCHED SAMPLES”. en. In: *Biometrika* 37.3-4 (Dec. 1950), pp. 256–266. ISSN: 0006-3444, 1464-3510. DOI: 10.1093/biomet/37.3-4.256. <https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/37.3-4.256> (visited on 04/04/2024).
- [22] Wilma Coetzee. “Measuring risks associated with students of introductory statistics: Scale development and implementation”. In: *Perspectives in Education* 40.2 (June 2022), pp. 143–158. ISSN: 2519-593X, 0258-2236. DOI: 10.18820/2519593X/pie.v40.i2.11. <https://journals.ufs.ac.za/index.php/pie/article/view/5643> (visited on 02/12/2025).
- [23] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. 2nd ed. New Jersey: Lawrence Erlbaum Associates, 1988. ISBN: 978-0-80-580283-2.
- [24] Jacob Cohen et al. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 3rd. Mahwah, NJ: Lawrence Erlbaum Associates, 2003.
- [25] Kathleen M. T. Collins and Anthony J. Onwuegbuzie. “I Cannot Read My Statistics Textbook: The Relationship Between Reading Ability and Statistics Anxiety”. In: *The Journal of Negro Education* 76.2 (2007). Publisher: Journal of Negro Education, pp. 118–129. ISSN: 00222984, 21676437. <http://www.jstor.org/stable/40034551> (visited on 04/08/2025).
- [26] W Conover and R Iman. *Multiple-comparisons procedures. Informal report*. Tech. rep. LA-7677-MS, 6057803. Feb. 1979, LA-7677-MS, 6057803. DOI: 10.2172/6057803. <https://www.osti.gov/servlets/purl/6057803/> (visited on 05/07/2024).
- [27] W. J. Conover. *Practical nonparametric statistics*. eng. 3rd ed. Wiley series in probability and statistics. Applied probability and statistics section. New York: Wiley, 1999. ISBN: 978-0-471-16068-7.
- [28] Harald Cramér. *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press, 1946.
- [29] R. J. Cruise, R. W. Cash, and D. L. Bolton. “Development and validation of an instrument to measure statistical anxiety”. In: *American Statistical Association 1985 Proceedings of the Section on Statistical Education*. Washington, D.C.: American Statistical Association, 1985, pp. 92–97.
- [30] Patrick Daleiden et al. “Analysis of a Randomized Controlled Trial of Student Performance in Parallel Programming using a New Measurement Technique”. In: *ACM Trans. Comput. Educ.* 20.3 (Sept. 2020). DOI: 10.1145/3401892. <https://doi-org.ezproxy.library.unlv.edu/10.1145/3401892>.

- [31] Carol I. Diener and Carol S. Dweck. “An analysis of learned helplessness: Continuous changes in performance, strategy, and achievement cognitions following failure.” en. In: *Journal of Personality and Social Psychology* 36.5 (May 1978), pp. 451–462. ISSN: 1939-1315, 0022-3514. DOI: 10.1037/0022-3514.36.5.451. <https://doi.apa.org/doi/10.1037/0022-3514.36.5.451> (visited on 06/05/2026).
- [32] Alexis Dinno. *Effect Size of Cochran’s Q Test*. Cross Validated (Stack Exchange). Accessed July 9, 2026. 2018. <https://stats.stackexchange.com/questions/9867/effect-size-of-cochrans-q>.
- [33] Yadolah Dodge. *The Concise Encyclopedia of Statistics*. en. New York, NY: Springer New York, 2008. ISBN: 978-0-387-31742-7 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1. <http://link.springer.com/10.1007/978-0-387-32833-1> (visited on 04/18/2025).
- [34] Olive Jean Dunn. “Multiple Comparisons Among Means”. In: *Journal of the American Statistical Association* 56.293 (1961), pp. 52–64. DOI: 10.1080/01621459.1961.10482090.
- [35] Olive Jean Dunn. “Multiple Comparisons Using Rank Sums”. In: *Technometrics* 6.3 (1964), pp. 241–252. DOI: 10.1080/00401706.1964.10490181.
- [36] Morgan Earp. “Development and Validation of the Statistics Anxiety Measure”. Unpublished Doctoral Dissertation. University of Denver, 2007.
- [37] M. L. Edirisooriya and T. J. Lipscomb. “Gender Influence on Statistics Anxiety among Graduate Students.” In: *Journal of Research in Science Mathematics and Technology Education* 4.2 (2021), pp. 63–74. <https://doi.org/10.31756/jrsmte.421>.
- [38] Allen L. Edwards. “Note on the ”correction for continuity” in testing the significance of the difference between correlated proportions”. en. In: *Psychometrika* 13.3 (Sept. 1948), pp. 185–187. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/BF02289261. <http://link.springer.com/10.1007/BF02289261> (visited on 04/04/2024).
- [39] Y. Eshet, P. Steinberger, and K. Grinautsky. “Does statistics anxiety impact academic dishonesty? Academic challenges in the age of distance learning.” In: *International Journal for Educational Integrity* 18.1 (2022), pp. 23–19. <https://doi.org/10.1007/s40979-022-00117-w>.
- [40] Yovav Eshet, Pnina Steinberger, and Keren Grinautsky. “Relationship between Statistics Anxiety and Academic Dishonesty: A Comparison between Learning Environments in Social Sciences”. en. In: *Sustainability* 13.3 (Feb. 2021), p. 1564. ISSN: 2071-1050. DOI: 10.3390/su13031564. <https://www.mdpi.com/2071-1050/13/3/1564> (visited on 04/08/2025).
- [41] Fernanda Fiel Peres. “Effect Sizes for Nonparametric Tests”. In: *Biochemia Medica* 36.1 (2025), p. 010101. DOI: 10.11613/BM.2026.010101.
- [42] Andy P. Field. *Discovering statistics using IBM SPSS statistics: and sex and drugs and rock ’n’ roll*. 4th edition. Los Angeles: Sage, 2013. ISBN: 978-1-4462-4917-8 978-1-4462-4918-5.

- [43] Andy P. Field, Jeremy Miles, and Zoë Field. *Discovering statistics using R*. OCLC: ocn760970657. London ; Thousand Oaks, Calif: Sage, 2012. ISBN: 978-1-4462-0046-9 978-1-4462-0045-2.
- [44] Sara J Finney and Gregory Schraw. “Self-efficacy beliefs in college statistics courses”. In: *Contemporary Educational Psychology* 28.2 (2003), pp. 161–186. ISSN: 0361-476X. DOI: [https://doi.org/10.1016/S0361-476X\(02\)00015-2](https://doi.org/10.1016/S0361-476X(02)00015-2). <https://www.sciencedirect.com/science/article/pii/S0361476X02000152>.
- [45] Ronald A. Fisher. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.
- [46] Paul M. Fitts. “The information capacity of the human motor system in controlling the amplitude of movement”. In: *Journal of Experimental Psychology* 47.6 (1954), pp. 381–391. DOI: 10.1037/h0055392.
- [47] C. Forrester et al. “Undergraduate R Programming Anxiety in Ecology: Persistent Gender Gaps and Coping Strategies”. In: *CBE Life Sciences Education* 21.2 (2022), pp. 1–18. DOI: 10.1187/cbe.21-05-0133. <https://doi.org/10.1187/cbe.21-05-0133>.
- [48] John Fox and Sanford Weisberg. *car: Companion to Applied Regression*. R package version 3.1-3. 2019. <https://CRAN.R-project.org/package=car>.
- [49] M. Frey-Clark, P. Natesan, and M. O’Bryant. “Assessing Statistical Anxiety Among Online and Traditional Students.” In: *Frontiers in Psychology* 10 (2019). <https://doi.org/10.3389/fpsyg.2019.01440>.
- [50] Milton Friedman. “The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance”. In: *Journal of the American Statistical Association* 32.200 (1937), pp. 675–701. DOI: 10.1080/01621459.1937.10503522.
- [51] Paul A. Games and John F. Howell. “Pairwise Multiple Comparison Procedures with Unequal N’s and/or Variances: A Monte Carlo Study”. In: *Journal of Educational and Behavioral Statistics* 1.2 (1976), pp. 113–125.
- [52] William W. Gaver. “Auditory icons: using sound in computer interfaces”. In: *Hum.-Comput. Interact.* 2.2 (June 1986), pp. 167–177. ISSN: 0737-0024. DOI: 10.1207/s15327051hci0202_3. https://doi.org/10.1207/s15327051hci0202_3.
- [53] F. Ghani and S. Maat. “Anxiety and Achievement in Statistics: A Systematic Review on Quantitative Studies.” In: *Creative Education* 9 (2018), pp. 2280–2290. DOI: 10.4236/ce.2018.914168.
- [54] S.N. Goodman, D.G. Altman, and S.L. George. “Statistical reviewing policies of medical journals.” In: *Journal of General Internal Medicine* 13 (1998), pp. 753–756. <https://doi.org/10.1046/j.1525-1497.1998.00227.x>.
- [55] Ali S. Hadi. *Regression Analysis by Example*. eng. 1st ed. New York Academy of Sciences Series. Newark: John Wiley & Sons, Incorporated, 2012. ISBN: 978-1-118-45624-8.

- [56] Donncha Hanna, Mark Shevlin, and Martin Dempster. “The structure of the statistics anxiety rating scale: A confirmatory factor analysis using UK psychology students”. In: *Personality and Individual Differences* 45.1 (2008), pp. 68–74. <https://doi.org/10.1016/j.paid.2008.02.021..>
- [57] T. E. Hardwicke and S. N. Goodman. “How often do leading biomedical journals use statistical experts to evaluate statistical methods? The results of a survey”. In: *PLoS ONE* 15.10 (2020), e0239598. DOI: 10.1371/journal.pone.0239598. <https://doi.org/10.1371/journal.pone.0239598>.
- [58] William L. Hays. *Statistics for Psychologists*. New York: Holt, Rinehart and Winston, 1963.
- [59] Sarah Hoegler and Mary Nelson. “The Influence of Anxiety and Self-Efficacy on Statistics Performance: A Path Analysis”. In: *Psi Chi Journal of Psychological Research* 23.5 (2018), pp. 364–375. ISSN: 23257342. DOI: 10.24839/2325-7342.JN23.5.364. <https://www.psichi.org/page/235JNWinter2018> (visited on 04/08/2025).
- [60] Harold Hotelling. “A Generalized T Test and Measure of Multivariate Dispersion”. In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1951, pp. 23–41.
- [61] T.-Y. Hsiao. “The Statistical Anxiety Rating Scale: Further Evidence for Multidimensionality.” In: *Psychological Reports* 107.3 (2010), pp. 977–982. <https://doi-org.ezproxy.library.unlv.edu/10.2466/07.11.PR0.107.6.977-982>.
- [62] Maxwell K. Hsu, Stephen W. Wang, and Kevin K. Chiu. “Computer attitude, statistics anxiety and self-efficacy on statistical software adoption behavior: An empirical study of online MBA learners”. en. In: *Computers in Human Behavior* 25.2 (Mar. 2009), pp. 412–420. ISSN: 07475632. DOI: 10.1016/j.chb.2008.10.003. <https://linkinghub.elsevier.com/retrieve/pii/S0747563208001866> (visited on 04/08/2025).
- [63] B. W. Hunt et al. “Statistics anxiety and predictions of exam performance in UK psychology students.” In: *PloS One* 18.8 (2023). <https://doi.org/10.1371/journal.pone.0290467>.
- [64] D. N. Joanes and C. A. Gill. “Comparing Measures of Sample Skewness and Kurtosis”. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1 (1998), pp. 183–189. DOI: 10.1111/1467-9884.00122.
- [65] Jibin Joseph. *Machine Learning to Predict Digital Frustration from Clickstream Data*. arXiv:2512.20438 [cs.LG]. Dec. 2025. DOI: 10.48550/arXiv.2512.20438. <http://arxiv.org/abs/2512.20438> (visited on 05/15/2026).
- [66] L. Kaufmann et al. “Self-efficacy matters: Influence of students’ perceived self-efficacy on statistics anxiety.” In: *Annals of the New York Academy of Sciences* 1514.1 (2022), pp. 187–197. <https://doi.org/10.1111/nyas.14797>.

- [67] Maurice G. Kendall. *Rank Correlation Methods*. 4th. London: Charles Griffin, 1970. ISBN: 978-0-85264-199-6.
- [68] Tan Hon Kian et al. “Anxiety and examination stress in statistics among postgraduate students”. eng. In: *AIP Conference Proceedings*. Vol. 2433. Melville: American Institute of Physics, 2022, p. 020030. ISBN: 9780735443907. DOI: 10.1063/5.0072652.
- [69] Young Ik Kim et al. “Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys”. In: *Social Science Computer Review* 37.2 (2019), pp. 214–233. DOI: 10.1177/0894439317752406.
- [70] J. Peter Kincaid, Richard Braby, and John E. Mears. “Electronic authoring and delivery of technical information”. en. In: *Journal of Instructional Development* 11.2 (June 1988), pp. 8–13. ISSN: 0162-2641. DOI: 10.1007/BF02904998. <http://link.springer.com/10.1007/BF02904998> (visited on 02/12/2025).
- [71] J. Peter Kincaid et al. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Tech. rep. Research Branch Report 8-75. Millington, TN: Naval Technical Training Command, Research Branch, 1975. <https://api.semanticscholar.org/CorpusID:61131325>.
- [72] Timothy Kluthe et al. “A randomized controlled trial on the nomenclature of scientific computing”. In: *Computer Science Education* 35.3 (2024), pp. 606–634. DOI: 10.1080/08993408.2024.2403971. eprint: <https://doi.org/10.1080/08993408.2024.2403971>. <https://doi.org/10.1080/08993408.2024.2403971>.
- [73] Clyde Y Kramer. “Extension of multiple range tests to group means with unequal numbers of replications”. In: *Biometrics* 12.3 (1956), pp. 307–310. DOI: 10.2307/3001469.
- [74] Jon A. Krosnick. “Response strategies for coping with the cognitive demands of attitude measures in surveys”. In: *Applied Cognitive Psychology* 5.3 (1991), pp. 213–236. DOI: <https://doi.org/10.1002/acp.2350050305>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350050305>. <https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.2350050305>.
- [75] William H. Kruskal and W. Allen Wallis. “Use of Ranks in One-Criterion Variance Analysis”. In: *Journal of the American Statistical Association* 47.260 (1952), pp. 583–621. DOI: 10.1080/01621459.1952.10483441.
- [76] J. Richard Landis and Gary G. Koch. “The Measurement of Observer Agreement for Categorical Data”. In: *Biometrics* 33.1 (Mar. 1977), p. 159. ISSN: 0006341X. DOI: 10.2307/2529310. <https://www.jstor.org/stable/2529310?origin=crossref> (visited on 01/22/2025).
- [77] Øyvind Langsrud. “ANOVA for Unbalanced Data: Use Type II Instead of Type III Sums of Squares”. In: *Statistics and Computing* 13.2 (2003), pp. 163–167. DOI: 10.1023/A:1023260610025.
- [78] D. N. Lawley. “A Generalization of Fisher’s z Test”. In: *Biometrika* 30.1/2 (1939), pp. 180–187.

- [79] Michael A. Lawrence. *ez: Easy Analysis and Visualization of Factorial Experiments*. R package version 3.0-0. 2011. <https://CRAN.R-project.org/package=ez>.
- [80] Carmen León-Mantero et al. “Analysis of attitudinal components towards statistics among students from different academic degrees”. en. In: *PLOS ONE* 15.1 (Jan. 2020). Ed. by Andrew R. Dalby, e0227213. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0227213. <https://dx.plos.org/10.1371/journal.pone.0227213> (visited on 02/12/2025).
- [81] Orly Lipka and Itay Hess. “Attitudes Toward Statistics Studies Among Students with Learning Disabilities.” In: *Numeracy* 9.2 (2016). <http://dx.doi.org/10.5038/1936-4660.9.2.7>.
- [82] Scott M. Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777. ISBN: 9781510860964.
- [83] P. Lytsy, M. Hartman, and R. Pingel. “Misinterpretations of P-values and statistical tests persists among researchers and professionals working with statistics and epidemiology”. In: *Upsala Journal of Medical Sciences* 127.1 (2022), e8760. DOI: 10.48101/ujms.v127.8760. <https://doi.org/10.48101/ujms.v127.8760>.
- [84] Kelly Rhea MacArthur and Jonathan B. Santo and. “A Multi-Level Analysis of the Effects of Statistics Anxiety/Attitudes on Trajectories of Exam Scores”. In: *Journal of Statistics and Data Science Education* 31.1 (2023), pp. 102–112. DOI: 10.1080/26939169.2022.2093805. eprint: <https://doi.org/10.1080/26939169.2022.2093805>. <https://doi.org/10.1080/26939169.2022.2093805>.
- [85] D. Macher et al. “Statistics anxiety and performance: blessings in disguise”. In: *Frontiers in Psychology* 6 (2015), p. 1116. DOI: 10.3389/fpsyg.2015.01116. <https://doi.org/10.3389/fpsyg.2015.01116>.
- [86] D. Macher et al. “Statistics anxiety, state anxiety during an examination, and academic achievement.” In: *The British Journal Of Educational Psychology* 83.4 (2013), pp. 535–549. <https://doi.org/10.1111/j.2044-8279.2012.02081.x>.
- [87] D. Macher et al. “Statistics anxiety, trait anxiety, learning behavior, and academic performance.” In: *European Journal of Psychology of Education* 27 (2012), pp. 483–498. <https://doi.org/10.1007/s10212-011-0090-5>.
- [88] Henry B. Mann and Donald R. Whitney. “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other”. In: *The Annals of Mathematical Statistics* 18.1 (1947), pp. 50–60. DOI: 10.1214/aoms/1177730491.
- [89] E. Masuadi et al. “Trends in the Usage of Statistical Software and Their Associated Study Designs in Health Sciences Research: A Bibliometric Analysis”. In: *Cureus* 13.1 (2021), e12639. DOI: 10.7759/cureus.12639. <https://doi.org/10.7759/cureus.12639>.

- [90] Nicholas J. McCaughey, Taylor G. Hill, and Sean P. Mackinnon. “The Association of Self-Efficacy, Anxiety Sensitivity, and Perfectionism With Statistics and Math Anxiety”. eng. In: *Personality Science* 3.1 (2022), e7091. ISSN: 2700-0710. DOI: 10.5964/ps.7091. <https://doi.org/10.5964/ps.7091>.
- [91] Quinn McNemar. “Note on the sampling error of the difference between correlated proportions or percentages”. en. In: *Psychometrika* 12.2 (June 1947), pp. 153–157. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/BF02295996. <http://link.springer.com/10.1007/BF02295996> (visited on 04/04/2024).
- [92] Renata A Mendes et al. “Statistics anxiety or statistics fear? A reinforcement sensitivity theory perspective on psychology students’ statistics anxiety, attitudes, and self-efficacy”. en. In: *Eur. J. Psychol. Educ.* 39.3 (Sept. 2024), pp. 2461–2480.
- [93] Ainsley Miller and Kate Pyper. “Anxiety Around Learning R in First Year Undergraduate Students: Mathematics versus Biomedical Sciences Students”. en. In: *Journal of Statistics and Data Science Education* 32.1 (Jan. 2024), pp. 47–53. ISSN: 2693-9169. DOI: 10.1080/26939169.2023.2190010. <https://www.tandfonline.com/doi/full/10.1080/26939169.2023.2190010> (visited on 02/12/2025).
- [94] Amanda Mireles. “Blended Pedagogy in Social Statistics Courses: Prelecture Strategies for Encouraging Learning among First-Generation College Students”. en. In: *Teaching Sociology* 51.3 (July 2023), pp. 231–244. ISSN: 0092-055X, 1939-862X. DOI: 10.1177/0092055X231170749. <https://journals.sagepub.com/doi/10.1177/0092055X231170749> (visited on 02/12/2025).
- [95] A. Mji. “Differences in university students’ attitudes and anxiety about statistics.” In: *Psychological Reports* 104.3 (2009), pp. 737–744. <https://doi.org/10.2466/PR0.104.3.737-744>.
- [96] J. A. Nelder and R. W. M. Wedderburn. “Generalized Linear Models”. In: *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), pp. 370–384. ISSN: 00359238, 23972327. <http://www.jstor.org/stable/2344614> (visited on 07/08/2026).
- [97] Peter Björn Nemenyi. “Distribution-free Multiple Comparisons”. PhD thesis. Princeton University, 1963.
- [98] Michael E. R. Nicholls et al. “Satisfaction guaranteed: The effect of spatial biases on responses to Likert scales”. In: *Psychological Science* 17.12 (2006), pp. 1027–1028. DOI: 10.1111/j.1467-9280.2006.01822.x.
- [99] Stephen Olejnik and James Algina. “Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs.” en. In: *Psychological Methods* 8.4 (2003), pp. 434–447. ISSN: 1939-1463, 1082-989X. DOI: 10.1037/1082-989X.8.4.434. <https://doi.apa.org/doi/10.1037/1082-989X.8.4.434> (visited on 04/30/2024).
- [100] Chester L. Olson. “On Choosing a Test Statistic in Multivariate Analysis of Variance”. In: *Psychological Bulletin* 83.4 (1976), pp. 579–586. DOI: 10.1037/0033-2909.83.4.579.

- [101] A. J. Onwuegbuzie and V. A. Wilson. “Statistics anxiety: Nature, etiology, antecedents, effects, and treatments: A comprehensive review of the literature.” In: *Teaching in Higher Education* 8 (2003), pp. 195–209. <https://doi.org/10.1080/1356251032000052447>.
- [102] Anthony J Onwuegbuzie and Christine E Daley. “Perfectionism and statistics anxiety”. en. In: *Personality and Individual Differences* 26.6 (June 1999), pp. 1089–1102. ISSN: 01918869. DOI: 10.1016/S0191-8869(98)00214-1. <https://linkinghub.elsevier.com/retrieve/pii/S0191886998002141> (visited on 04/02/2025).
- [103] Anthony J. Onwuegbuzie. “Academic procrastination and statistics anxiety”. en. In: *Assessment & Evaluation in Higher Education* 29.1 (Feb. 2004), pp. 3–19. ISSN: 0260-2938, 1469-297X. DOI: 10.1080/0260293042000160384. <http://www.tandfonline.com/doi/abs/10.1080/0260293042000160384> (visited on 02/12/2025).
- [104] Anthony J. Onwuegbuzie. “Attitudes toward Statistics Assessments”. en. In: *Assessment & Evaluation in Higher Education* 25.4 (Dec. 2000), pp. 321–339. ISSN: 0260-2938, 1469-297X. DOI: 10.1080/713611437. <http://www.tandfonline.com/doi/abs/10.1080/713611437> (visited on 02/12/2025).
- [105] Anthony J. Onwuegbuzie and Kathleen M. T. Collins. “Reading Comprehension among Graduate Students”. en. In: *Psychological Reports* 90.3 (June 2002), pp. 879–882. ISSN: 0033-2941, 1558-691X. DOI: 10.2466/pr0.2002.90.3.879. <https://journals.sagepub.com/doi/10.2466/pr0.2002.90.3.879> (visited on 04/08/2025).
- [106] Karl Pearson. “Note on Regression and Inheritance in the Case of Two Parents”. In: *Proceedings of the Royal Society of London* 58 (1895), pp. 240–242. DOI: 10.1098/rspl.1895.0041.
- [107] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [108] Á. Peiró-Signes et al. “Anxiety towards Statistics and Its Relationship with Students’ Attitudes and Learning Approach”. In: *Behavioral Sciences (Basel, Switzerland)* 11.3 (2021), p. 32. DOI: 10.3390/bs11030032. <https://doi.org/10.3390/bs11030032>.
- [109] Christopher Peterson, Steven F Maier, and Martin E P Seligman. *Learned Helplessness: A Theory for the Age of Personal Control*. en. Oxford University Press New York, NY, July 1993. ISBN: 978-0-19-504466-9 978-0-19-773596-1. DOI: 10.1093/oso/9780195044669.001.0001. <https://academic.oup.com/book/53580> (visited on 06/05/2026).
- [110] K. C. S. Pillai. “Some New Test Criteria in Multivariate Analysis”. In: *Annals of Mathematical Statistics* 26.1 (1955), pp. 117–121.

- [111] Timothy Rafalski et al. “A Randomized Controlled Trial on the Wild Wild West of Scientific Computing with Student Learners”. In: *Proceedings of the 2019 ACM Conference on International Computing Education Research*. ICER '19. Toronto ON, Canada: Association for Computing Machinery, 2019, pp. 239–247. ISBN: 9781450361859. DOI: 10.1145/3291279.3339421. <https://doi-org.ezproxy.library.unlv.edu/10.1145/3291279.3339421>.
- [112] Kevin M. Ralston et al. “Anxious women or complacent men? Anxiety of statistics in a sample of UK sociology undergraduates”. In: *International Journal of Social Research Methodology* 24 (2020), pp. 79–91. <https://api.semanticscholar.org/CorpusID:219486715>.
- [113] Kevin M. Ralston et al. *We need to talk about statistical anxiety. A review of the evidence around statistical anxiety in the context of quantitative methods pedagogy*. NCRM Working Paper. National Centre for Research Methods (NCRM), 2016. <https://api.semanticscholar.org/CorpusID:151353772>.
- [114] P. Rendulic and S. Terrell. “Anxiety toward statistics and the use of computers: A study of graduate level education majors.” In: *Journal of Computing in Higher Education* 11.2 (2012), pp. 104–120. <https://doi.org/10.1007/BF02940893>.
- [115] J. B. Rode and M. M. Ringel. “Statistical Software Output in the Classroom: A Comparison of R and SPSS”. In: *Teaching of Psychology* 46.4 (2019), pp. 319–327. DOI: 10.1177/0098628319872605. <https://doi.org/10.1177/0098628319872605>.
- [116] Ma. Mercedes T. Rodrigo and Ryan S.J.d. Baker. “Coarse-grained detection of student frustration in an introductory programming course”. In: *Proceedings of the Fifth International Workshop on Computing Education Research Workshop*. ICER '09. Berkeley, CA, USA: Association for Computing Machinery, 2009, pp. 75–80. ISBN: 9781605586151. DOI: 10.1145/1584322.1584332. <https://doi.org/10.1145/1584322.1584332>.
- [117] Neil Salkind. *Encyclopedia of Measurement and Statistics*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: Sage Publications, Inc., 2007. ISBN: 978-1-4129-1611-0 978-1-4129-5264-4. DOI: 10.4135/9781412952644. <https://methods.sagepub.com/reference/encyclopedia-of-measurement-and-statistics> (visited on 04/18/2025).
- [118] Kate Sanders et al. “Inferential Statistics in Computing Education Research: A Methodological Review”. In: *Proceedings of the 2019 ACM Conference on International Computing Education Research*. ICER '19. Toronto ON, Canada: Association for Computing Machinery, 2019, pp. 177–185. ISBN: 9781450361859. DOI: 10.1145/3291279.3339408. <https://doi.org/10.1145/3291279.3339408>.
- [119] Emily K. Sandoz, Grayson Butcher, and Tracy A. Protti. “A preliminary examination of willingness and importance as moderators of the relationship between statistics anxiety and performance”. en. In: *Journal of Contextual Behavioral Science* 6.1 (Jan. 2017), pp. 47–52. ISSN: 22121447. DOI: 10.1016/j.jcbs.2017.02.002. <https://linkinghub.elsevier.com/retrieve/pii/S2212144717300054> (visited on 04/08/2025).

- [120] Candace Schau et al. “The Development and Validation of the Survey of Attitudes toward Statistics”. en. In: *Educational and Psychological Measurement* 55.5 (Oct. 1995), pp. 868–875. ISSN: 0013-1644, 1552-3888. DOI: 10.1177/0013164495055005022. <http://journals.sagepub.com/doi/10.1177/0013164495055005022> (visited on 03/26/2024).
- [121] C. James Scheirer, William S. Ray, and N. Hare. “The Analysis of Ranked Data Derived from Completely Randomized Factorial Designs”. In: *Biometrics* 32.2 (1976), pp. 429–434. DOI: 10.2307/2529511.
- [122] Jocelyn Scheirer et al. “Frustrating the user on purpose: a step toward building an affective computer”. In: *Interacting with Computers* 14.2 (2002), pp. 93–118. ISSN: 0953-5438. DOI: [https://doi.org/10.1016/S0953-5438\(01\)00059-5](https://doi.org/10.1016/S0953-5438(01)00059-5). <https://www.sciencedirect.com/science/article/pii/S0953543801000595>.
- [123] Shayle R. Searle, F. M. Speed, and George A. Milliken. “Population Marginal Means in the Linear Model: An Alternative to Least Squares Means”. In: *The American Statistician* 34.4 (1980), pp. 216–221. DOI: 10.1080/00031305.1980.10483031.
- [124] M E P Seligman. “Learned Helplessness”. en. In: *Annual Review of Medicine* 23.1 (Feb. 1972), pp. 407–412. ISSN: 0066-4219, 1545-326X. DOI: 10.1146/annurev.me.23.020172.002203. <https://www.annualreviews.org/doi/10.1146/annurev.me.23.020172.002203> (visited on 05/29/2026).
- [125] Carrie E. Smith and Robert A. Cribbie. “Factorial ANOVA with Unbalanced Data: A Fresh Look at the Types of Sums of Squares”. In: *Journal of Data Science* 12.2 (2014), pp. 385–404.
- [126] Eric van Sonderen, Robbert Sanderman, and James C. Coyne. “Ineffectiveness of Reverse Wording of Questionnaire Items: Let’s Learn from Cows in the Rain”. In: *PLOS ONE* 8.7 (July 2013), pp. 1–7. DOI: 10.1371/journal.pone.0068967. <https://doi.org/10.1371/journal.pone.0068967>.
- [127] Charles Spearman. “The Proof and Measurement of Association Between Two Things”. In: *The American Journal of Psychology* 15.1 (1904), pp. 72–101. DOI: 10.2307/1412159.
- [128] “Spearman Rank Correlation Coefficient”. en. In: *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, 2008, pp. 502–505. ISBN: 978-0-387-31742-7 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1_379. http://link.springer.com/10.1007/978-0-387-32833-1_379 (visited on 04/02/2024).
- [129] Student. “The Probable Error of a Mean”. In: *Biometrika* 6.1 (1908), pp. 1–25. DOI: 10.1093/biomet/6.1.1.
- [130] Javier Suárez-Álvarez et al. “Using reversed items in Likert scales: A questionable practice”. In: *Psicothema* 30.2 (2018), pp. 149–158. DOI: 10.7334/psicothema2018.33.
- [131] Scott D. Swain, Danny Weathers, and Ronald W. Niedrich. “Assessing Three Sources of Misresponse to Reversed Likert Items”. In: *Journal of Marketing Research* 45.1 (2008), pp. 116–131. DOI: 10.1509/jmkr.45.1.116. eprint: <https://journals.sagepub.com/doi/pdf/10.1509/jmkr.45.1.116>. <https://journals.sagepub.com/doi/abs/10.1509/jmkr.45.1.116>.

- [132] John Sweller. “Cognitive Load During Problem Solving: Effects on Learning”. In: *Cognitive Science* 12.2 (1988), pp. 257–285. DOI: https://doi.org/10.1207/s15516709cog1202_4. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1202_4. https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1202_4.
- [133] John Sweller, Jeroen J. G. Van Merriënboer, and Fred G. W. C. Paas. “Cognitive Architecture and Instructional Design”. en. In: *Educational Psychology Review* 10.3 (Sept. 1998), pp. 251–296. ISSN: 1040-726X, 1573-336X. DOI: 10.1023/A:1022193728205. <https://link.springer.com/10.1023/A:1022193728205> (visited on 05/29/2026).
- [134] AP. Trassi et al. “Mediating factors of statistics anxiety in university students: a systematic review and meta-analysis.” In: *Annals of the New York Academy of Sciences* 1512.1 (June 2022), pp. 76–97. DOI: 10.1111/nyas.14746.
- [135] John W. Tukey. “Comparing Individual Means in the Analysis of Variance”. In: *Biometrics* 5.2 (1949), pp. 99–114.
- [136] Wade M. Vagias. *Likert-type scale response anchors*. Tech. rep. Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management, Clemson University, 2006.
- [137] Raphael Vallat. “Pingouin: statistics in Python”. In: *Journal of Open Source Software* 3.31 (Nov. 2018), p. 1026. DOI: 10.21105/joss.01026.
- [138] Sharon Vaughn and Marcia A. Barnes. “Reading comprehension for students with reading disabilities: Progress and challenges”. en. In: *Learning and Individual Differences* 102 (Feb. 2023), p. 102258. ISSN: 10416080. DOI: 10.1016/j.lindif.2023.102258. <https://linkinghub.elsevier.com/retrieve/pii/S1041608023000031> (visited on 02/12/2025).
- [139] Andreu Vigil-Colet, Urbano Lorenzo-Seva, and Lorena Condon. “Development and validation of the statistical anxiety scale.” In: *Psicothema* 20.1 (2008), pp. 174–80.
- [140] A. Von Mayrhauser and A.M. Vans. “Program comprehension during software maintenance and evolution”. In: *Computer* 28.8 (1995), pp. 44–55. DOI: 10.1109/2.402076.
- [141] W3C Web Accessibility Initiative. *Web Content Accessibility Guidelines (WCAG)*. Tech. rep. W3C Recommendation. World Wide Web Consortium (W3C), 2025. <https://www.w3.org/WAI/standards-guidelines/wcag/> (visited on 08/23/2025).
- [142] Julie Waples. “Statistics Anxiety and Statistics Self-Efficacy in Doctoral Online Statistics: A Mixed Methods Study”. en. In: (2020). Publisher: Unpublished. DOI: 10.13140/RG.2.2.13976.83207. <http://rgdoi.net/10.13140/RG.2.2.13976.83207> (visited on 04/02/2025).

- [143] B. L. Welch. “On the Comparison of Several Mean Values: An Alternative Approach”. In: *Biometrika* 38.3/4 (Dec. 1951), p. 330. ISSN: 00063444. DOI: 10.2307/2332579. <https://www.jstor.org/stable/2332579?origin=crossref> (visited on 04/30/2024).
- [144] B. L. Welch. “The Generalization of Student’s Problem When Several Different Population Variances Are Involved”. In: *Biometrika* 34.1/2 (1947), pp. 28–35. DOI: 10.2307/2332510.
- [145] Hadley Wickham. “Tidy Data”. In: *Journal of Statistical Software* 59.10 (2014), pp. 1–23. DOI: 10.18637/jss.v059.i10. <https://www.jstatsoft.org/index.php/jss/article/view/v059i10>.
- [146] David Gray Widder, Joshua Sunshine, and Stephen Fickas. “Barriers to Reproducible Scientific Programming”. In: *2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. Memphis, TN, USA: IEEE, 2019, pp. 217–221. DOI: 10.1109/VLHCC.2019.8818907.
- [147] Frank Wilcoxon. “Individual Comparisons by Ranking Methods”. In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83. DOI: 10.2307/3001968.
- [148] G. N. Wilkinson and C. E. Rogers. “Symbolic Description of Factorial Models for Analysis of Variance”. In: *Applied Statistics* 22.3 (1973), p. 392. ISSN: 00359254. DOI: 10.2307/2346786. <https://www.jstor.org/stable/10.2307/2346786?origin=crossref> (visited on 06/09/2026).
- [149] Samuel S. Wilks. “Certain Generalizations in the Analysis of Variance”. In: *Biometrika* 24.3/4 (1932), pp. 471–494.
- [150] Hannah Williams, Timothy Kluthe, and Andreas Stefik. “Does Learning Statistics Have to Be So Difficult? Evidence from a Randomized Controlled Trial”. In: *ACM Trans. Comput. Educ.* (July 2026). Just Accepted. DOI: 10.1145/3831705. <https://doi.org/10.1145/3831705>.
- [151] Christa E. Winkler. “Advanced skills in quantitative assessment”. In: *New Directions for Student Services* 2022.178-179 (2022), pp. 133–148. DOI: <https://doi.org/10.1002/ss.20435>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ss.20435>. <https://onlinelibrary.wiley.com/doi/abs/10.1002/ss.20435>.
- [152] R. Wu et al. “Reducing the influence of perfectionism and statistics anxiety on college student performance in statistics courses.” In: *Frontiers In Psychology* 13 (2022). <https://doi.org/10.3389/fpsyg.2022.1011278>.
- [153] Moshe Zeidner. “STATISTICS AND MATHEMATICS ANXIETY IN SOCIAL SCIENCE STUDENTS: SOME INTERESTING PARALLELS”. en. In: *British Journal of Educational Psychology* 61.3 (Nov. 1991), pp. 319–328. ISSN: 0007-0998, 2044-8279. DOI: 10.1111/j.2044-8279.1991.tb00989.x. <https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/j.2044-8279.1991.tb00989.x> (visited on 04/09/2025).

Curriculum Vitae

Graduate College
University of Nevada, Las Vegas

Hannah C. Williams

hannah.c.williams11@gmail.com

Degrees:

Bachelor of Science – Computer Science
University of Nevada, Las Vegas
May 2022

Thesis Title: Unveiling the Digital Age of Data Science: An Evidence-Driven Investigation

Thesis Examination Committee:

Chairperson, Dr. Andreas Stefik, Ph.D.
Committee Member, Dr. Fatma Nasoz, Ph.D.
Committee Member, Dr. Laxmi Gewali, Ph.D.
Committee Member, Dr. Hal Berghel, Ph.D.
Graduate Faculty Representative, Dr. Sarah Harris, Ph.D.