



Evaluating the Usability of Jamovi

For One-Way ANOVA Analysis

LI Kanyu 

1. OVERVIEW

1.a. Problem Statement and Context

Statistical software interfaces function as the primary interaction point between researchers and data, making their design a crucial factor in preventing analytical mistakes. Jamovi has positioned itself as an accessible alternative to traditional tools, yet the question remains: does its user experience design provide the necessary guidance to ensure methodological rigor, or does it prioritize ease of use at the expense of analytical precision?

One-way Analysis of Variance (ANOVA) represents an ideal test case for usability evaluation: it requires sufficient analytical complexity to expose interface weaknesses while remaining accessible to researchers across disciplines. The complete workflow, data import, cleaning, assumption checking, analysis execution, and interpretation, creates multiple points where poor interface design can induce user error.

This research addresses potential gap in understanding how statistical software design impacts research quality. Given Jamovi’s positioning as a free, accessible alternative to commercial platforms, understanding its usability has implications for research reproducibility, educational contexts, and open science initiatives.

1.b. Objectives

Primary Research Question:

To what extent does Jamovi’s user interface facilitate or hinder the accurate completion of one-way ANOVA analysis workflows?

Secondary Research Questions:

- What specific usability issues emerge at different stages of the statistical analysis workflow (data import, cleaning, analysis, visualization, interpretation)?
- How do usability perceptions and problem discovery differ between general STEM users and UX-trained evaluators?
- What are the capabilities and limitations of AI-assisted usability evaluation compared to human-centered methods?

1.c. Research Outputs

This progressive research plan will produce three deliverable reports:

Usability Study Report (Due: February 10, 2026)

- Task performance metrics (completion time, error rates, success rates)
- Standardized usability measurements (SUS, NASA-TLX)
- Qualitative insights from think-aloud protocols and interviews

Drafted Jan 18, 2026

Contents

Overview	1
Methods	2
Research Timeline	10
Budget	11
Ethical Compliance	12
Appendices	14
References	17
AI usage declaration	18

Correspondence to

LI Kanyu
im@not.ci

Competing Interests

The authors declare no competing interests.

PURE Study Report (Due: March 10, 2026)

- All outputs in Study 1
- Detailed step-by-step discoverability and ease-of-use ratings
- Expert heuristic evaluation findings with severity ratings
- Comparative analysis of expert vs. novice problem discovery

AI Evaluation Study Report (Due: March 24, 2026)

- Automated usability testing results across 30 independent trials
- AI-guided heuristic evaluations and affinity mapping
- Comparative synthesis across all three methodologies
- Critical assessment of AI's role in UX research workflows

2. METHODS*2.a. User-Centered Usability Testing**2.a.i. Method Selection and Justification:*

This study employs **mixed-methods usability testing** with concurrent think-aloud protocol (Ericsson & Simon, 1993). This approach is justified by the fact that:

- **Industry standard:** User testing is the industry's mainstream standard method involving real users (Nielsen, 1995)
- **Think-aloud validity:** Concurrent verbalization effectively captures cognitive processes during task execution without significantly altering performance (Ericsson, 2006)
- **Sample size rationale:** Five participants can uncover approximately 85% of usability problems for interfaces of moderate complexity (Nielsen & Landauer, 1993)

*2.a.ii. Participants (n=5):***Inclusion Criteria:**

- STEM disciplinary background (science, technology, engineering, mathematics)
- Completed at least one university-level statistics course
- Basic understanding of basic concepts of frequentist inference
- No formal UX training or professional UX experience
- Able to participate via Zoom with reliable internet connection

Exclusion Criteria:

- Current UX students or professionals
- Extensive prior Jamovi experience (>5 hours of use)
- Statistical software developers

Recruitment Strategy:

- Recruitment via social media networks and personal/professional contacts
- Screening based on educational background and professional experience
- Target diversity in STEM disciplines (mix of natural sciences, engineering, health sciences)

2.a.iii. Testing Protocol:

Pre-Test Phase:

- Welcome and rapport building
- Review and sign informed consent form (electronic signature)
- Brief explanation of think-aloud protocol with demonstration
- Equipment check (Zoom recording, screen sharing, audio quality)

Task Execution Phase:

Participants assume the role of a sleep researcher investigating the question: “Does eating cheese before bed cause nightmares?” They complete four sequential tasks using Jamovi:

Task 1: Data Import and cleaning

Instruction: “Import the file cheese_nightmare_data.csv into Jamovi and verify the data loaded correctly. Then examine the dataset for errors or anomalies. Clean the data as appropriate for analysis.”

Hidden challenges: Dataset contains one extreme outlier (data entry error) and one format error (letter in numeric field)

Success criteria: File imported; participant confirms data structure; participant identifies and addresses data quality issues

Task 2: One-Way ANOVA Execution

Instruction: “Conduct a one-way ANOVA comparing nightmare frequency across different cheese types. Provide statistical interpretation of your findings.”

Success criteria: Correct ANOVA specification; appropriate interpretation of results

Task 3: Data Visualization

Instruction: “Create an appropriate chart that clearly displays the relationship between cheese type and nightmare occurrence.”

Success criteria: Suitable graph type selected; proper formatting for interpretation

Post-Task Assessment:

- System Usability Scale (SUS) - 10 items
- NASA Task Load Index (NASA-TLX) - 6 dimensions

Semi-Structured Interview:

Follow-up questions exploring user experience:

1. “Overall, how would you describe your experience using Jamovi for this analysis?”
2. “Which task was easiest? Why?”
3. “Which task was most difficult or frustrating? What specifically made it challenging?”
4. “Were there moments when you felt confused or stuck? What was happening at that point?”
5. “Were there times when you felt uncertain about what to do next? Can you describe those situations?”
6. [Probe specific observed behaviors:] “I noticed you [specific action]. Can you tell me what you were thinking at that moment?”

Closing:

- Thank participant for their time

- Provide compensation/acknowledgment
- Stop recording and save files

2.a.iv. *Data Collection Instruments:*

Quantitative Metrics:

1. **Task Performance:**
 - Completion time per task (timestamp from instruction delivery to participant declaration of completion)
 - Task success rate (binary: complete/incomplete)
 - Error count (number of incorrect actions or backtracking instances)
 - Subjective task difficulty experience (5 points Likert)
2. **Standardized Assessments:**
 - **System Usability Scale (SUS):** 10-item questionnaire; scores range 0-100; threshold for acceptable usability: > 7 (Brooke, 1996)
 - **NASA Task Load Index (NASA-TLX):** 6-dimension workload assessment; threshold for high workload: >54 (Hart & Staveland, 1988)

Qualitative Data:

1. **Think-Aloud Transcripts:** Audio recordings transcribed verbatim
2. **Semi-Structured Interview Data:** Coded for recurring themes, pain points, and mental models
3. **Observer Notes:** Including critical incidents (errors, confusion, recovery attempts), Emotional responses (frustration, satisfaction, surprise), navigation patterns and workflow strategies, unsolicited comments or questions

Testing Environment:

- Remote participation via Zoom (ecological validity through participant's own environment)
- Participants use personal computers (Windows/Mac compatibility check during recruitment)
- Jamovi pre-installed by participant with setup verification
- Screen recording and audio capture via Zoom cloud recording

Test Materials Provided:

- Informed consent form (PDF)
- cheese_nightmare_data.csv (100 simulated observations: cheese_type [4 levels: Cheddar, Brie, Swiss, Blue], nightmare_frequency [continuous])
- SUS questionnaire (Microsoft Forms)
- NASA-TLX questionnaire (Microsoft Forms)

2.a.v. *Data Analysis Plan:*

Quantitative Analysis:

- Descriptive statistics for completion times, error rates, SUS scores, NASA-TLX dimensions
- Individual task performance profiles
- Correlation analysis between subjective ratings and objective performance

Qualitative Analysis:

- **AI-Assisted Coding**
 - Gemini 3.0 Pro ([Pichai et al., 2025](#)) processes all transcripts 30 times independently
 - Compare thematic consistency across runs
 - Final themes represent convergent findings across iterations and build codebook from the output
 - Convert Qualitative codec into Quantitative code to build linear regression model to isolate factors that affect user experience
- **Affinity Mapping:**
 - Extract key user quotes and observations
 - Group related issues into thematic clusters
 - Identify patterns in pain points, confusion, and success moments
 - Create visual affinity diagram linking user needs to interface failures

2.b. Study 2: Pragmatic Usability Rating by Experts (PURE)

2.b.i. Method Selection and Justification:

The PURE methodology bridges traditional expert evaluation and user testing by having UX professionals complete actual tasks while applying expert heuristics ([Rohrer et al., 2016](#)). This approach is justified by:

- **Theory-Practice Integration:** Experts experience firsthand usability issues while applying theoretical frameworks
- **Problem Discovery Enhancement:** Trained evaluators can identify issues that novice users may normalize or fail to articulate
- **Cost-Effectiveness:** 3-5 expert evaluators can discover 60-75% of usability issues with significantly lower resource requirements than extensive user testing ([Fu et al., 2002](#); [Nielsen & Molich, 1990](#))

2.b.ii. Evaluators (n=3):

Inclusion Criteria:

- Current enrollment in UX Design graduate program (minimum one semester completed)
- STEM undergraduate background (psychology, cognitive science, or related fields preferred)
- Completed university-level statistics course
- Familiar with heuristic evaluation methodology
- No professional relationship with Jamovi development

Exclusion Criteria:

- Participants from Study 1 (user testing)
- Jamovi development team members or contributors

Recruitment Strategy:

- Recruit from UX604 cohort (Winter 2026)
- Target students with quantitative research backgrounds (psychology, HCI, data science)
- Emphasis on independent assessment

2.b.iii. Testing Protocol:

Phase 1: Task Completion with Think-Aloud

Evaluators complete identical tasks as Study 1 participants:

- Task 1: Data Import and Validation and Cleaning
- Task 2: One-Way ANOVA Execution
- Task 3: Data Visualization

Concurrent verbalization captures expert mental models and immediate reactions.

Phase 2: Step-by-Step PURE Evaluation

Evaluators review a standardized Standard Operating Procedure (SOP) document breaking the workflow into granular steps (approximately 25-30 steps total). For each step, evaluators rate:

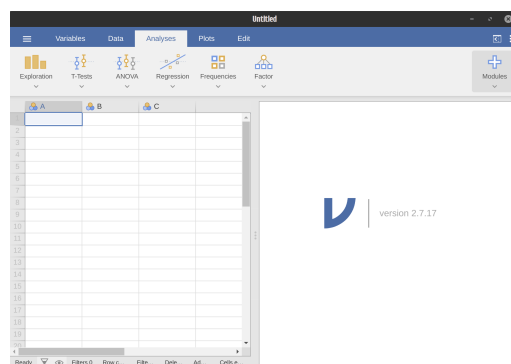
Discoverability (5-point scale):

- 5 = Immediately obvious, clear visual cues
- 4 = Easy to find with minimal exploration
- 3 = Requires moderate searching or menu navigation
- 2 = Hidden, requires significant effort to locate
- 1 = Nearly impossible to discover without external guidance

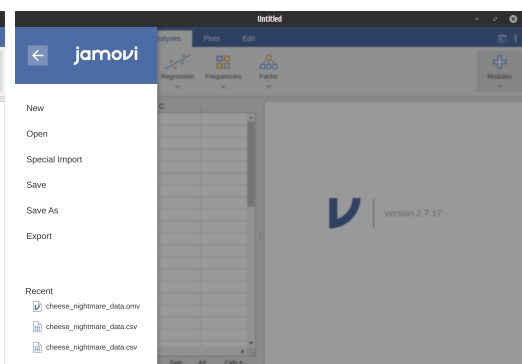
Ease of Use (5-point scale):

- 5 = Intuitive, requires no prior knowledge
- 4 = Easy with minimal learning curve
- 3 = Manageable with some trial-and-error
- 2 = Difficult, frustrating, error-prone
- 1 = Extremely challenging, likely to cause errors

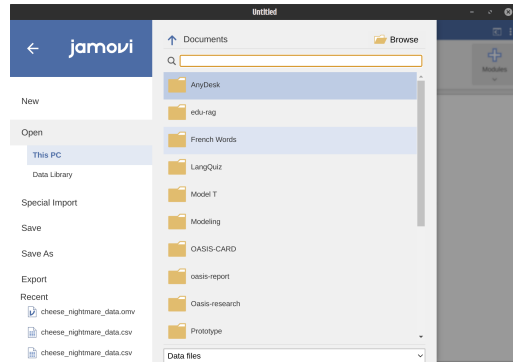
Task SOPs:



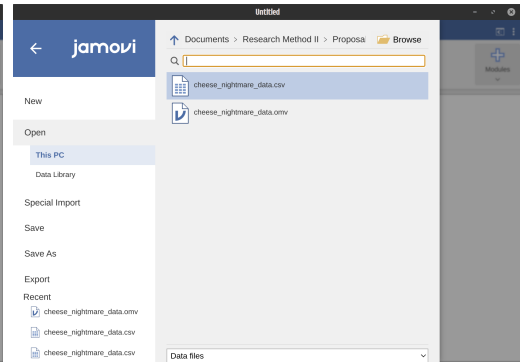
Open "Jamovi"



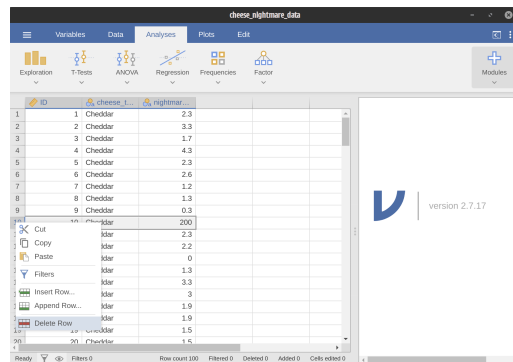
Address the File Menu



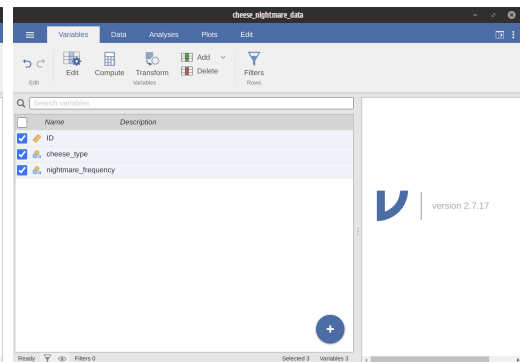
Locate the "Open" function in Jamovi



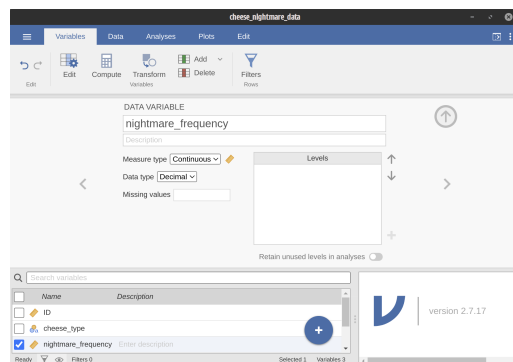
Find the file in the directory



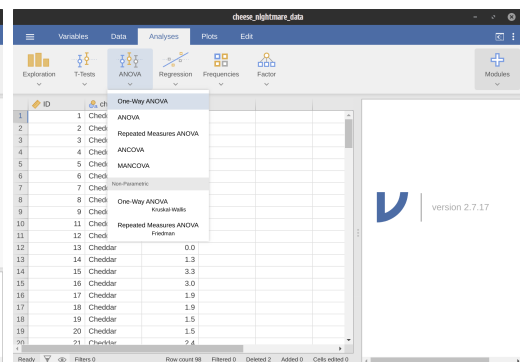
Identify the dirty data and remove them



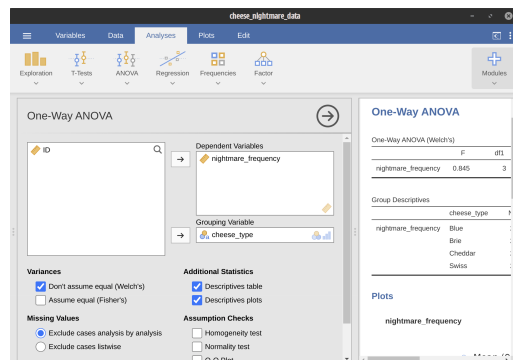
Identify the data type interface



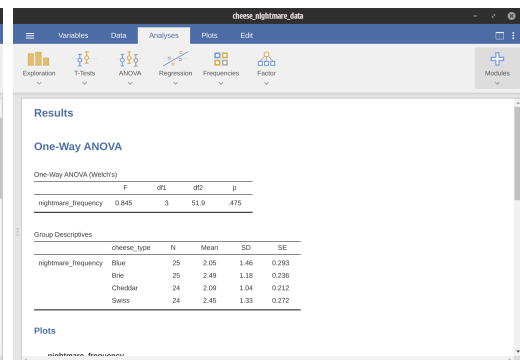
Correct the data type



Access the ANOVA analysis menu



Specify variables and functions



Interpret the ANOVA output table

Phase 3: Heuristic Evaluation

Independent evaluation against Nielsen's 10 Usability Heuristics:

1. **Visibility of System Status**
2. **Match Between System and Real World**
3. **User Control and Freedom**
4. **Consistency and Standards**
5. **Error Prevention**
6. **Recognition Rather Than Recall**
7. **Flexibility and Efficiency of Use**
8. **Aesthetic and Minimalist Design**
9. **Help Users Recognize, Diagnose, and Recover from Errors**
10. **Help and Documentation**

For each usability issue identified, evaluators document:

- **Description:** Clear statement of the problem
- **Location:** Specific interface element or workflow stage
- **Heuristic Violation(s):** Which principle(s) are violated
- **Severity Rating:**
 - 0 = Not a usability problem
 - 1 = Cosmetic issue (fix if time permits)
 - 2 = Minor usability problem (low priority)
 - 3 = Major usability problem (high priority)
 - 4 = Usability catastrophe (must fix before release)
- **Recommended Fix:** Specific, actionable improvement suggestion

Phase 4: Post-Evaluation

- SUS questionnaire
- NASA-TLX assessment
- Brief reflection on evaluation process

2.b.iv. Data Collection:

Quantitative Data:

- Task performance metrics (identical to Study 1)
- PURE ratings: Discoverability and Ease of Use scores for each SOP step
- Heuristic violation counts by category
- Severity distribution of identified issues
- SUS and NASA-TLX scores

Qualitative Data:

- Think-aloud transcripts during task execution
- Detailed heuristic evaluation reports
- Recommended improvements with rationale
- Observer notes on expert strategies

2.b.v. *Data Analysis Plan:*

Quantitative Analysis:

- Aggregate PURE ratings across evaluators (mean, SD, inter-rater reliability)
- Identify workflow steps with lowest discoverability/ease scores
- Frequency analysis of heuristic violations
- Severity-weighted problem prioritization
- Comparative analysis: Expert vs. novice performance metrics

Qualitative Analysis:

- Thematic analysis of expert think-alouds (compare to Study 1 novice patterns)
- Content analysis of heuristic evaluation reports
- Categorization of recommended improvements by impact and feasibility
- Synthesis of expert mental models and problem-solving strategies

Cross-Study Synthesis:

- Issues identified only by experts vs. only by users vs. both
- Alignment between user-reported frustrations and heuristic violations
- Expert prediction accuracy for novice struggles

2.c. *Study 3: AI-Assisted Usability Evaluation*

2.c.i. *Method Selection and Justification:*

This study explores emerging capabilities of AI agents in usability assessment through:

1. **Automated Task Execution:** Gemini 2.5 (Comanici et al., 2025) Computer Use model independently completes identical tasks 30 times
2. **Computational Thematic Analysis:** LLMs synthesize qualitative data from Studies 1 and 2

Rationale:

- **Scalability:** AI enables large-scale independent replications impossible with human participants, generating robust statistical distributions
- **Consistency Control:** Each trial begins with cleared context, ensuring independence and comparability

2.c.ii. *AI Tools and Configuration:*

Task Execution Prompt Template:

You are a sleep researcher investigating cheese consumption and nightmares. You have access to a computer with Jamovi statistical software installed.

Your research question: "Does eating cheese before bed cause nightmares?"

The data is located in {{PATH}}, your job is load and clean the data, conduct an one way ANOVA with proper visualization then give me your conclusion.

Thematic Analysis Prompt Template:

You are a qualitative research analyst. Analyze the following usability test transcripts to identify recurring themes, pain points, and patterns.

{{TRANSCRIPT}}

YOUR TASK:

1. Identify common themes across participants related to:
 - Confusion or difficulty points
 - Frustration or dissatisfaction
 - Moments of success or ease
 - Unmet expectations
 - Suggestions for improvement
2. For each theme:
 - Provide a clear theme name and description
 - List supporting quotes with participant IDs (e.g., P02, E01)
 - Count how many participants mentioned this theme
 - Note which task stages the theme relates to
 - Assess severity/importance

Focus on themes that appear across multiple participants and that have clear implications for interface redesign.

2.c.iii. Data Analysis Plan:

Automated Task Performance:

- Success rate distribution across 30 trials (binomial test for each task)
- Completion time descriptive statistics
- Error pattern frequency analysis
- Comparison to human performance (Studies 1 & 2)

Thematic Analysis Validation:

- Convergence analysis: theme stability across 30 independent LLM runs
- Compare AI-generated themes to researcher's manual coding
- Identify novel insights vs. confirmation of human findings
- Evaluate AI's capacity for nuanced qualitative interpretation

3. RESEARCH TIMELINE

3.a. Detailed Schedule

Phase	Dates	Deliverables	Time
Research Plan Development	Jan 6-20	• Finalize methodology • Prepare test materials • Develop consent forms &	2 weeks

		questionnaires • Submit Research Plan	
Study 1: User Testing			
Participant recruitment	Jan 21-27	• Screen 8-10 candidates • Confirm 5 participants • Schedule sessions	1 week
Data collection	Jan 28-Feb 3	• Conduct 5 usability sessions (1/day) • Observer notes & recordings	1 week
Analysis & reporting	Feb 4-10	• Transcribe sessions • Thematic analysis • Calculate metrics • Submit Usability Study Report	1 week
Study 2: PURE Evaluation			
Evaluator recruitment	Feb 18-23	• Recruit 3 UX students • Verify STEM backgrounds • Schedule sessions	<1 week
Data collection	Feb 24-Mar 2	• Conduct 3 expert evaluations • Heuristic analysis sessions	1 week
Analysis & reporting	Mar 3-10	• Aggregate PURE ratings • Synthesize heuristic findings • Comparative analysis with Study 1 • Submit PURE Study Report	1 week
Study 3: AI Evaluation			
AI infrastructure setup	Mar 11-12	• Configure Gemini API • Test virtual environment • Validate prompts	2 days
Automated data collection	Mar 13-16	• Run 30 AI task trials • Run 30 heuristic evaluations • Run 30 thematic analysis iterations	4 days
Analysis & synthesis	Mar 17-24	• Statistical analysis of AI performance • Validate AI-generated themes • Cross-method comparison • Final synthesis & recommendations • Submit AI Evaluation Report	1 week

3.b. Risk Mitigation

- **Backup participants:** 3 additional candidates in the pool for backup purpose
- **Equipment failure:** Zoom alternative (Google Meet) prepared

4. BUDGET

Category	Item	Cost	Justification
Software	Jamovi	\$0	Open-source, freely available
	Zoom	\$0	Using existing university license

Category	Item	Cost	Justification
	Gemini API	\$0	Free tier sufficient for 30 trials via AI Studio
	Microsoft Forms	\$0	Included in university account
	Transcription	\$0	Using Gemini Audio (free tier)
Participant	User testing	\$0	Friends/family volunteer basis
	PURE evaluators	\$0	Classmate exchange
Materials	Test dataset	\$0	Researcher-generated simulated data
	Consent forms	\$0	Digital templates (digital.gov)
	Questionnaires	\$0	Standard validated scales (public domain)
Equipment	Researcher computer	\$0	Personal device
	Participants' computers	\$0	Bring your own device (BYOD)
Total		\$0	

5. ETHICAL COMPLIANCE

5.a. TCPS 2 Compliance

This research adheres to the **Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2)**, specifically:

Core Principles:

1. **Respect for Persons:** Voluntary informed consent; right to withdraw
2. **Concern for Welfare:** Minimal risk; protecting participant wellbeing
3. **Justice:** Fair participant selection; equitable burden/benefit distribution

Rationale for No REB Submission:

- Research involves minimal risk (software usability testing)
- No vulnerable populations targeted
- No collection of sensitive personal information
- No deception or psychological manipulation
- Educational context (course assignment)
- Qualifies for delegated review / exemption under TCPS 2 Article 2.5 ([Canadian Institutes of Health Research et al., 2022](#))

However, research still follows REB-level standards:

- Full informed consent process
- Data protection and anonymization protocols
- Right to withdraw without consequences
- Secure data storage and destruction plan

5.b. Data Protection Plan

Anonymization:

- Participant IDs assigned upon consent (P01-P05, E01-E03)
- Master list linking IDs to identities kept in separate file
- All data files, transcripts, reports use IDs only

- Demographic information reported in aggregate only (no individual identification possible)

Storage Security:

- Recordings stored on cloud storage (pCloud) with 2-factor authentication and encryption
- No data stored on public or shared computers

Access Control:

- Only researcher has access to identifiable data
- Course instructor may review anonymized data for grading
- No third-party access without explicit participant consent

Data Retention and Destruction:

- Active use: January 20 - March 24, 2026 (course duration)
- Retention: Until April 15, 2026 (2 weeks post-course for grading)
- Destruction: Permanent deletion of all recordings, transcripts, and identifiable information by April 15, 2026
- Anonymized summary data may be retained for portfolio use (no identifiers)

5.c. Participant Rights

Voluntary Participation:

- Participation entirely voluntary
- No penalty for declining or withdrawing
- No impact on relationship with university or researcher

Right to Withdraw:

- Participants may withdraw at any time before, during, or up to 1 week after their session
- Withdrawal results in complete data deletion
- No explanation required

Right to Skip:

- Participants may decline to answer any question
- Participants may decline to complete any task
- Partial participation acceptable

Debriefing:

- Participants receive summary of research findings upon request (after March 24, 2026)
- Contact information provided for questions or concerns
- Resources for participant support if any distress experienced (unlikely but available)

5.d. Ethical Considerations for AI Use

Transparency:

- Full disclosure that AI tools are used for evaluation
- Clear documentation of prompts, model versions, and parameters
- Acknowledgment of AI's role in analysis sections of reports

Bias Awareness:

- AI reflects training data biases; findings interpreted with this caveat

- Cross-validation with human-generated analyses required
- AI used to augment, not replace, human judgment

Data Privacy:

- No personally identifiable information from human participants included in AI prompts
- Transcripts anonymized before AI processing
- API usage complies with Google's terms of service

Academic Integrity:

- AI-generated content clearly attributed
- Critical analysis of AI outputs (not blindly accepted)
- AI serves as research tool, not substitute for researcher thinking

6. APPENDICES

6.a. Test Materials Checklist

- ☐ cheese_nightmare_data.csv (100 rows, variables: cheese_type, nightmare_frequency)
- ☐ Task instruction sheet (PDF)
- ☐ Informed consent form
- ☐ SUS questionnaire (Microsoft Form)
- ☐ NASA-TLX questionnaire (Microsoft Form)
- ☐ Semi-structured interview guide
- ☐ Observer notes template
- ☐ Participant screening questionnaire
- ☐ PURE Standard Operating Procedure document (30 steps)
- ☐ Heuristic evaluation reporting template

6.b. Appendix B: Participant Tracking Templates

User Testing Tracker (Study 1): [Spreadsheet with columns: ID, Recruitment Date, Screening Score, STEM Field, Session Date/Time, Consent Signed, Recording Saved, Transcription Complete, SUS Score, NASA-TLX Score, Notes]

PURE Evaluator Tracker (Study 2): [Spreadsheet with columns: ID, UX Courses Completed, STEM Background, Session Date/Time, PURE Ratings Complete, Heuristic Eval Complete, Total Issues Identified, Severity Distribution]

AI Trial Tracker (Study 3): [Spreadsheet with columns: Trial ID, Timestamp, Task 1 Success, Task 2 Success, Task 3 Success, Task 4 Success, Total Time, Errors Logged, Notes]

6.c. Sample Consent Form

Thank you for participating in our usability research. We will be recording your session to allow LI Kanyu, who is not present, to benefit from your comments.

Please read the statement below and sign where indicated.

I understand that my usability test session will be recorded.

I grant LI Kanyu permission to use this recording for internal use only, for the purpose of improving the designs being tested.

I understand and accept that my think-aloud protocols and interview responses conducted during this experiment may be transcribed and anonymously quoted in academic papers.

Signature: _____

Date: _____

6.d. Screening Questionnaire

1. What is your highest level of education completed?
 - ☐ Bachelor's degree
 - ☐ Master's degree
 - ☐ PhD
 - ☐ Other: _____
2. What is your field of study/work? (open text)
3. Have you completed a university-level statistics course?
 - ☐ Yes → Which course(s)? _____
 - ☐ No → [Exclude]
4. Rate your familiarity with these statistical concepts (1=Never heard of it, 5=Very comfortable):
 - Hypothesis testing: 1 2 3 4 5
 - P-values: 1 2 3 4 5
 - Analysis of Variance (ANOVA): 1 2 3 4 5

[Minimum threshold: Average > 3]
5. Do you have professional experience in UX design or research?
 - ☐ Yes → [Exclude]
 - ☐ No
6. How many hours have you used Jamovi statistical software?
 - ☐ 0 hours
 - ☐ 1-5 hours

- ☐ 6-10 hours → [Exclude]
 - ☐ 10+ hours → [Exclude]
7. Are you available for a 1-hour Zoom session during [date range]?
- ☐ Yes → Preferred times: _____
 - ☐ No
8. Do you have access to:
- ☐ Computer (Windows/Mac) with admin rights to install software
 - ☐ Stable internet connection for video call
 - ☐ Quiet space for 1 hour

Selection Criteria:

- First 5 qualified respondents meeting all criteria
- Backup pool of 3 additional qualified candidates in case of cancellations

6.e. Mock Data Generation Code

```
set.seed(114)

n_per_group <- 25
groups <- c("Cheddar", "Brie", "Swiss", "Blue")
total_n <- n_per_group * length(groups)

cheese_type <- rep(groups, each = n_per_group)

freq_cheddar <- rnorm(n_per_group, mean = 2.1, sd = 1)
freq_brie <- rnorm(n_per_group, mean = 2.5, sd = 1.2)
freq_swiss <- rnorm(n_per_group, mean = 2, sd = 1.3)
freq_blue <- rnorm(n_per_group, mean = 2.3, sd = 1.5)

nightmare_frequency <- c(freq_cheddar, freq_brie, freq_swiss,
freq_blue)

nightmare_frequency <- pmax(nightmare_frequency, 0)
nightmare_frequency <- round(nightmare_frequency, 1)

df <- data.frame(
  ID = 1:total_n,
  cheese_type = factor(cheese_type, levels = groups),
  nightmare_frequency = as.character(nightmare_frequency),
  stringsAsFactors = FALSE
)

outlier_row <- 10
df$nightmare_frequency[outlier_row] <- "200"
print(paste("Injected outlier at row:", outlier_row, "Value: 200"))
```

```

format_error_row <- 60
df$nightmare_frequency[format_error_row] <- "3a"
print(paste("Injected format error at row:", format_error_row, "Value:
'3a'"))

write.csv(df, "cheese_nightmare_data.csv", row.names = FALSE, quote =
FALSE)

```

REFERENCES

- Anthropic. (2024,). *Claude 4.5 Sonnet*. <https://www.anthropic.com/claude>
- Brooke, e., John. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, 189(194), 4–7.
- Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, & Social Sciences and Humanities Research Council of Canada. (2022). *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 2022*. <https://tcps2core.ca/welcome>
- Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N.-J., ... Helmholtz, W. (2025). Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *Arxiv Preprint Arxiv:2507.06261*. <https://arxiv.org/abs/2507.06261>
- Ericsson, K. A. (2006). Protocol Analysis and Expert Thought: Concurrent Verbalizations of Thinking during Experts' Performance on Representative Tasks. In *The Cambridge Handbook of Expertise and Expert Performance* (pp. 223–242). Cambridge University Press. <https://doi.org/10.1017/cbo9780511816796.013>
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*. The MIT Press. <https://doi.org/10.7551/mitpress/5657.001.0001>
- Fu, L., Salvendy, G., & Turley, L. (2002). Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour & Information Technology*, 21(2), 137–143. <https://doi.org/10.1080/02699050110113688>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology: Vol. 52. Advances in psychology* (pp. 139–183). Elsevier.
- Nielsen, J. (1995). Usability inspection methods. *Conference Companion on Human Factors in Computing Systems - CHI '95*, 377–378. <https://doi.org/10.1145/223355.223730>
- Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '93*, 206–213. <https://doi.org/10.1145/169059.169166>
- Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Empowering People - CHI '90*, 249–256. <https://doi.org/10.1145/97243.97281>
- Pichai, S., Hassabis, D., & Kavukcuoglu, K. (2025, November). *A new era of intelligence with Gemini 3*. <https://blog.google/products-and-platforms/products/gemini/gemini-3/>
- Rohrer, C. P., Wendt, J., Sauro, J., Boyle, F., & Cole, S. (2016). Practical Usability Rating by Experts (PURE): A Pragmatic Approach for Scoring Product Usability. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 786–795. <https://doi.org/10.1145/2851581.2851607>

7. AI USAGE DECLARATION

This document is written with the help of Claude 4.5 ([Anthropic, 2024](#)).

Assignment rubric and syllable was provided as attachment.

Original	Translated (with Google Translate)
这是一份作业要求，我会把脑袋里面的东西全都倾倒出来，请你按照 Rubric 检查一下有无缺失，并向我提问。在确认信息周全后，你会整理一份完整的底稿，我会在此底稿的基础之上进行进一步的修改。 研究问题: Jamovi 的用户体验研究，以单因素 ANOVA 为例。 为什么是这个研究。相当多统计软件都遵循和 SPSS 类似的界面设计逻辑，Jamovi 也是其中的一员。但是以我个人的使用经验来看，这个设计范式有很多可以改进的空间以尽最大可能减少研究人员的误操作，得出错误的统计结论。单因素方差分析作为一个具有一定复杂度但又不过分的分析工具，是切入此研究问题的合宜入口。而 Jamovi 作为一个开源、免费、使用体验相对优秀的平台，是一个实施开放、可复现研究的优秀目标和基线。它的流程涉及到完整的数据导入、数据清理、类型检查、统计操作和结果解读。每一步都有可能因为系统的「自动化」、不良的设计指引导致用户出现错误。本研究希望识别主要的可用性问题和已有的人本设计方法论进行交叉比对，并探索使用 AI 工具进行用户体验研究的潜在可能性。 本研究会包含基本的用户可用性研究，寻找五名具备理工农医药背景的专业人员完成基本的问卷、访谈和实验任务，接下来，我们会寻找三名具备此类背景的 UX 专业人士完成相同的任务，但是引入和 UX 紧密相关的评估方法(启发式评估)。此外，在独立完成任务之后，我们也会列出标准操作步骤，要求实验参与者系统性的对每一个步骤进行评估。	This is an assignment requirement. I will pour out everything I know, and please check it against Rubric for any omissions and ask me any questions. After confirming that all information is complete, you will compile a complete draft, which I will then revise further. Research Question: A user experience study of Jamovi, using one-way ANOVA as an example. Why this research? Many statistical software programs follow a similar interface design logic to SPSS, and Jamovi is one of them. However, based on my personal experience, this design paradigm has much room for improvement to minimize researcher errors and erroneous statistical conclusions. One-way ANOVA, as an analytical tool with a certain level of complexity but not excessive complexity, is a suitable entry point for this research question. Jamovi, as an open-source, free platform with a relatively excellent user experience, is an excellent target and baseline for conducting open and reproducible research. Its process involves complete data import, data cleaning, type checking, statistical operations, and result interpretation. Each step is potentially prone to user errors due to the system's "automation" or poor design guidance. This study aims to identify key usability issues, cross-reference them with existing human-centered design methodologies, and explore the potential of using AI tools for user experience research.

最后，AI 评估部分会根据当下情况使用 Claude Computer Use、OpenAI Operator、Gemini Computer Use 中的一个或多个模型，独立完成数次任务，计算每一步的成功率和总成功率。同时，AI 也会根据过往提供的数据完成 Affinity Map 和启发式评估以提供更加多元的视角。

在本研究中，我们要主要回答的研究唯二你是：Jamovi 的用户体验酒精是帮助了还是伤害了统计分析任务？在比较不同专业水平的用户、UX 领域专家、和 AI 完成相同任务时，有哪些新的洞见能够浮现出来？

在本研究中，我们希望识别在完成单因素方差分析时主要的可用性问题，比较不同水平（自评）的用户、UX 领域专家和 AI 在行为模式上的差异。识别不同统计分析阶段 Jamovi 给用户带来的认知负荷为何？

实验设计方面，本实验会采用被试内设计，即每个被试完成所有受测任务。实验参与者会扮演一名睡眠领域的研究专家，通过单因素分析回答「睡前吃起司会不会导致噩梦」这个问题。这个研究题目被刻意设计成中性的状态（即背景知识不会影响任务执行，只有统计素养会），它们会使用 Jamovi 分别完成数据导入、数据清洗、执行单因素方差分析、可视化分析结果、给出结论这几步。在研究过程中，受试者被要求「出声思考」，以便于让我们知道实验参与者的思维脉络。全程会有录音录像，实验开始之前会签署知情同意书，知情同意书使用美国政府网站上提供的标准模板 (<https://digital.gov/resources/digitalgov-user-experience-resources/digitalgov-user-experience-program-usability-starter-kit>)。

本研究会使用混合研究方法，同时包含质化数据和量化数据结果。质化数据包含研究访谈、启发式评估结果，而量化数据则是标准问卷结果。

This study will include basic user usability research, recruiting five professionals with backgrounds in science, engineering, agriculture, and medicine to complete basic questionnaires, interviews, and experimental tasks. Next, we will recruit three UX professionals with similar backgrounds to complete the same tasks, but introduce evaluation methods closely related to UX (heuristic evaluation). Furthermore, after independently completing the tasks, we will outline standard operating procedures, requiring participants to systematically evaluate each step.

Finally, the AI evaluation will use one or more models from Claude Computer Use, OpenAI Operator, and Gemini Computer Use, depending on the current situation, to independently complete the task several times, calculating the success rate of each step and the overall success rate. Simultaneously, the AI will also perform Affinity Mapping and heuristic evaluation based on previously provided data to provide more diverse perspectives.

In this study, we primarily aim to answer two research questions: Did Jamovi's user experience aid or harm the statistical analysis task? What new insights emerge when comparing users at different skill levels, UX experts, and AI performing the same task?

In this study, we aim to identify the main usability issues in completing one-way ANOVA and compare the differences in behavioral patterns among users at different skill levels (self-reported), UX experts, and AI. We also aim to identify the cognitive load that Jamovi imposes on users at different stages of statistical analysis.

In terms of experimental design, this experiment will employ a within-subjects design, meaning each participant will

用户测试是行业「主流」、涉及「真实用户」的标准方法（1994 年发表的 *Usability Inspection Methods* 这篇论文有提到这个论点），而「大声思考」是获取关于潜在思维过程的有效数据（*Protocol Analysis and Expert Thought: Concurrent Verbalizations of Thinking during Experts' Performance on Representative Tasks* K. Anders Ericsson），而对于平均难度的操作界面五名用户足够揭示 85% 的可用性问题的（*A mathematical model of the finding of usability problems* 这篇文章提到了这点）。

实验部分，我们主要设计了几个任务：

任务一导入数据并验证：将数据导入 `cheese_nightmare_data.csv`，并且检查数据导入正确。任务二数据清理：发现数据当中的错误和异常，将它们清理掉。任务三执行单因素方差分析：请进行一个单因素方差分析，比较吃不同品种的起司，夜晚做噩梦的发生频次，并且给出统计见解。任务四可视化：绘制一个得当的图标，用来展示你的发现。这个图表需要清晰地展示食用起司的品种和夜晚做噩梦之间的关系。

我所关心的指标：任务执行效率（完成任务所需要的时间），在视频上打点计时，以研究者给完指示为起点、实验参与者出声表示任务完成为终点。任务执行有效性：成功与否、中间出现失误的次数系统可用性：SUS 分数（做个标准问卷，十题）、NASA-TLX 问卷（六题）

我所采集的质性数据：出声思考的语音转档半结构化的访谈观察笔记（遵循 digital.gov 上的实用文档：https://s3.amazonaws.com/digitalgov/_legacy-img/2014/01/instructions-for-usability-test-observers.doc）

实验设备：因为征召的实验参与者会通过 Zoom 会议的方式完成任务，所以

complete all tested tasks. Participants will assume the role of a sleep research expert, answering the question “Does eating cheese before bed cause nightmares?” through one-way ANOVA. This research question is intentionally designed to be neutral (i.e., background knowledge will not affect task performance; only statistical literacy will). They will use Jamovi to complete the steps of data import, data cleaning, performing one-way ANOVA, visualizing the analysis results, and providing conclusions. During the study, participants are asked to “think aloud” to allow us to understand their thought process. The entire process will be audio and video recorded. Informed consent will be obtained before the experiment begins, using the standard template provided on the US government website (<https://digital.gov/resources/digitalgov-user-experience-resources/digitalgov-user-experience-program-usability-starter-kit>).

This study will use a mixed research methodology, incorporating both qualitative and quantitative data. Qualitative data includes research interviews and heuristic assessments, while quantitative data consists of standard questionnaire results.

User testing is a mainstream, standard method in the industry involving real users (this point is mentioned in the 1994 paper *Usability Inspection Methods*), and “thinking aloud” is an effective way to obtain data on underlying thought processes (*Protocol Analysis and Expert Thought: Concurrent Verbalizations of Thinking during Experts' Performance on Representative Tasks*, K. Anders Ericsson). For an interface of average difficulty, five users are sufficient to reveal 85% of usability issues (this is mentioned in the paper *A mathematical model of the finding of usability problems*).

用户使用自己舒适的设备完成任务（高生态效度）。用户安装 Jamovi 软件，录音录像通过 Zoom 内置功能完成。 准备的实验材料包括：测试数据集 cheese_nightmare_data.csv (100 条模拟数据，包含是否做噩梦和睡前吃的起司品种) 任务指导书 (PDF) SUS 问卷, NASA-TLX 问卷 (Microsoft Form) 半结构化访谈提纲观察笔记模板 (PDF) 知情同意书 (PDF 电子签名) 测试环境实验参与者被要求寻找一个安静的环境完成任务，运行 Jamovi 全程观察不指导大概需要一小时完成所有任务。 任务设计：介绍（五分钟）：欢迎、签同意书、解释什么是大声思考任务执行（二十分钟到半小时）：完成四个任务问卷（五分钟）半结构化访谈（十分钟）寒暄，总结：感谢、停止录影。 对于 PURE 实验结构没差很多，但是在完成实验之后，专家需要对着标准操作步骤进行评分，评价它们的可发现性和易用性。并且，在这之后进行启发式评估。 PURE 方法将专家理论知识融合到了可用性测试当中 (Practical Usability Rating by Experts(PURE): A Pragmatic Approach for Scoring Product Usability 这篇论文有提到)。专家评估者能识别新手用户可能习以为常或难以言表的问题。采用 3 名评估者既能确保全面性，又符合资源限制 (3-5 名专家是一个比较有性价比的数字，Heuristic evaluation of user interfaces By Nielsen 和 Effectiveness of user testing and heuristic evaluation as a function of performance classification 这两篇内容可以支持这个内容)。 这些研究方法可以比较一般用户和专业用户的思维模式是否有差异，也可以帮助我们发觉更多从专业用户的角度对 UX 设计的思考。	In the experimental section, we designed several tasks: Task 1: Import and validate data: Import the data into cheese_nightmare_data.csv and verify that the data import is correct. Task 2: Data cleaning: Identify and remove errors and anomalies from the data. Task 3: Perform one-way ANOVA: Conduct a one-way ANOVA to compare the frequency of nightmares at night with the consumption of different types of cheese, and provide statistical insights. Task 4: Visualization: Create an appropriate chart to illustrate your findings. This chart should clearly show the relationship between the type of cheese consumed and nightmares at night. The metrics I am interested in: Task performance efficiency (time required to complete the task), timed by ticking dots on the video, starting from when the researcher gives instructions and ending when the participant verbally indicates task completion. Task Execution Effectiveness: Success or failure, number of errors encountered. System Usability: SUS score (standard questionnaire, 10 questions), NASA-TLX questionnaire (6 questions). Qualitative Data Collected: Audio recordings of spoken thought; Semi-structured interviews; Observation notes (following the practical documentation on digital.gov: https://s3.amazonaws.com/digitalgov/_legacy-img/2014/01/instructions-for-usability-test-observers.doc). Experimental Equipment: Since recruited participants would complete the task via Zoom meetings, users used their
--	---

研究三: AI 分析 AI 会通过 Computer Use 独立执行此任务三十次以形成近似正态的分布, 并对成功率进行假设检验, 同时, AI 被用于针对所有访谈记录进行统合, 进行启发性分析, 同时独立制作 Affinity Map 与我的手工分析形成比较和互补。 同样的, 为了确保数据的稳定性, 此测评会独立执行三十次, 绘制分布, 最后再由 LLM 合成成一个稳定的结论。 AI 在这里扮演很重要的角色, 因为我们可以轻易地通过新 session 清理上下文完成大量独立测评, 形成稳定、基于大量数据的、可控的分布, 便于统计分析。此外, 随着人工智能日益融入用户体验工作流程, 评估其在可用性评估中的能力具有实践与理论价值。这既契合研究者构建人工智能原生数据分析工具的目标, 也为理解人工智能在用户体验评估中的优势与局限提供了洞见。 日程计划方面: 三个实验基本上都是一周收数据两周写子实验报告 (包括数据分析和报告撰写, 这是一个相对比较合理的数字) 最后再用一周把所有结果拼合在一起。 考虑到都是亲友测验和同班同学测验, 所以没有 budget。 用的软件都是开源的, 服务也都是免费的。Google Gemini 有免费 API 额度, 整体上整个研究应该是不需要花什么钱的。	own comfortable devices (high ecological validity). Users installed Jamovi software, and audio and video recording were completed using Zoom's built-in functions. The prepared experimental materials include: Test dataset cheese_nightmare_data.csv (100 simulated data points, including whether nightmares occurred and the type of cheese eaten before bed) Task instructions (PDF) SUS questionnaire, NASA-TLX questionnaire (Microsoft Form) Semi-structured interview outline Observation note template (PDF) Informed consent form (PDF with electronic signature) Test Environment Participants are required to find a quiet environment to complete the tasks and run Jamovi. The entire process is observed without guidance. It takes approximately one hour to complete all tasks. Task Design: Introduction (5 minutes): Welcome, signing the consent form, explaining what thinking aloud is Task Execution (20 to 30 minutes): Completing four tasks Questionnaire (5 minutes) Semi-structured Interview (10 minutes) Friendly Remarks, Summary: Thank you, stop recording. The structure is not much different from the PURE experiment, but after the experiment, experts need to rate the standard operating procedures, evaluating their discoverability and ease of use. Furthermore, a heuristic evaluation is conducted afterward. The PURE method integrates expert theoretical knowledge into usability testing (mentioned in the paper "Practical Usability Rating by Experts (PURE): A Pragmatic Approach for Scoring Product Usability"). Expert evaluators can identify problems that novice users might take for granted or find difficult to articulate. Using three evaluators ensures
--	--

	comprehensiveness while adhering to resource constraints (3-5 experts is a cost-effective number, supported by the papers “Heuristic evaluation of user interfaces by Nielsen” and “Effectiveness of user testing and heuristic evaluation as a function of performance classification”). These research methods can compare the thinking patterns of general users and professional users, and help us discover more insights into UX design from a professional user’s perspective. Study 3: AI Analysis AI independently performed this task thirty times using Computer Use to form an approximately normal distribution and performed hypothesis testing on the success rate. Simultaneously, AI was used to synthesize all interview records for heuristic analysis and independently create an Affinity Map to compare and complement my manual analysis. Similarly, to ensure data stability, this evaluation will be run independently thirty times, the distribution will be plotted, and finally, LLM will be used to synthesize a stable conclusion. AI plays a crucial role here because we can easily complete a large number of independent evaluations by cleaning the context with new sessions, forming a stable, large-scale, and controllable distribution that facilitates statistical analysis. Furthermore, as AI is increasingly integrated into user experience workflows, assessing its capabilities in usability evaluation has both practical and theoretical value. This aligns with researchers’ goals of building AI-native data analysis tools and provides insights into understanding the advantages and limitations of AI in user experience evaluation. Regarding the timeline:
--	--

	The three experiments will generally involve one week for data collection, two weeks for writing sub-experiment reports (including data analysis and report writing, which is a relatively reasonable schedule), and finally one week for combining all results. Considering that the tests are conducted by friends, family, and classmates, there is no budget. The software used is all open-source, and the services are all free. Google Gemini offers free API credits, so overall, the research should not require much financial expenditure.
我已经仔细阅读了你的研究构想。整体框架非常清晰且有理论支撑，但在正式撰写前，我需要向你确认一些细节，以确保符合 Rubric 的所有要求： 关键缺失信息 1. 参与者招募细节 ● 用户测试组（5 人）： ▶ 具体的筛选标准是什么？（例如：统计学背景要求、Jamovi 使用经验要求） ▶ 筛选问卷的具体问题有哪些？ ▶ 招募渠道？（朋友圈、专业社群？） ▶ 如何确保他们有理工农医药背景但不是 UX 专家？ ● PURE 专家组（3 人）： ▶ UX 专业资质的具体要求？（多少年经验？是否需要相关认证？） ▶ 是否也需要统计学背景？ ▶ 如何避免与用户测试组重叠？ 2. 伦理审查 ● Rubric 明确要求 “participant consent” 和 “ethical protocol” ● 你提到使用 digital.gov 的模板，但需要说明： ▶ 是否需要通过学校的 Research Ethics Board (REB) 审批？	I have carefully read your research proposal. The overall framework is very clear and theoretically sound, but before formally writing it, I need to confirm some details with you to ensure it meets all Rubric’s requirements: Key Missing Information 1. Participant Recruitment Details ● User Testing Group (5 people): ● What are the specific screening criteria? (e.g., statistics background required, Jamovi experience required) ● What are the specific questions in the screening questionnaire? ● Recruitment channels? (Friends and social media, professional communities?) ● How to ensure they have a background in STEM fields but are not UX experts? ● PURE Expert Group (3 people): ● What are the specific requirements for UX professional qualifications? (How many years of experience? Are relevant certifications required?) ● Is a statistics background also required? ● How to avoid overlap with the user testing group? 2. Ethical Review

- ▶ 知情同意书是否包含：数据保存期限、退出机制、匿名化处理方式？
- ▶ 录音录像的存储与销毁计划？

3. AI 工具选择的最终决定

- 你提到“Claude Computer Use、OpenAI Operator、Gemini Computer Use 中的一个或多个”
- 需要确定：
 - ▶ 最终使用哪一个？还是三个都用进行对比？
 - ▶ 如果用多个，如何设计对比实验？
 - ▶ 30 次独立执行是每个工具各 30 次，还是总共 30 次？

4. 数据分析计划

- 量化数据：
 - ▶ 任务完成时间、错误次数的具体分析方法？（描述性统计？组间比较用什么检验？）
 - ▶ SUS 和 NASA-TLX 的阈值设定？（多少分算可接受？）
- 质性数据：
 - ▶ 访谈转录后的编码框架？（主题分析？扎根理论？）
 - ▶ Affinity Map 的具体制作流程？
 - ▶ 如何确保编码的信度？（是否需要第二编码员？）

5. 时间线的具体日期 Rubric 要求 “show schedule with dates, times, deliverables”，你目前只给了周数。需要：

- 具体的起止日期（例如：1 月 X 日-1 月 X 日）
- 每个里程碑的 deliverable 是什么？
- 是否有 contingency plan（如果招募延迟怎么办？）

6. 研究目标的优先级 你提到多个研究问题（RQ），需要明确：

- 主要 RQ 是什么？次要 RQ 是什么？
 - 三个研究方法分别主要回答哪个 RQ？
- 补充确认问题

A. 测试材料的完整性

- Rubric explicitly requires “participant consent” and “ethical protocol”
- You mentioned using the digital.gov template, but please clarify:
- Is approval from the university’s Research Ethics Board (REB) required?
- Does the informed consent form include: data retention period, exit mechanism, and anonymization method?
- What is the storage and destruction plan for audio and video recordings?

3. Final Decision on AI Tool Selection

- You mentioned “one or more of Claude Computer Use, OpenAI Operator, and Gemini Computer Use”
- Please determine:
- Which tool will be used ultimately? Or will all three be used for comparison?
- If multiple tools are used, how will the comparative experiments be designed?
- Does the 30 independent executions consist of 30 executions for each tool, or a total of 30 executions?

4. Data Analysis Plan

- **Quantifying Data:**
- What are the specific analysis methods for task completion time and error count? (Descriptive statistics? What tests are used for between-group comparisons?)
- Threshold settings for SUS and NASA-TLX? (What score is acceptable?)
- **Qualitative data:**
- Coding framework after interview transcription? (Topic analysis? Grounded theory?)
- Specific workflow for creating the Affinity Map?
- How to ensure coding reliability? (Is a second coder needed?)

5. Specific dates in the timeline

Rubrik requires “show schedule with dates, times, deliverables,” but you have only provided the number of weeks. You need:

● cheese_nightmare_data.csv 中的“错误和异常”具体是什么？（缺失值？离群值？格式错误？） ● 任务指导书是否包含成功标准？（例如：“正确识别出 X 个异常值”） B. 半结构化访谈提纲 能否列出 3-5 个核心问题？例如： 1) 哪个步骤最让你困惑？ 2) 你期望软件如何改进？ 3) … C. PURE 评分标准 ● “可发现性”和“易用性”的评分量表是什么？（1-5 分？ 1-7 分？） ● 标准操作步骤（SOP）的粒度如何？（总共多少步？） D. 启发式评估的具体执行 ● 10 条启发式原则是否全部使用？ ● 严重性评分（0-4）的定义？ ● 专家是独立评估还是集体讨论？ 我的初步建议 优势保持： ● ☒ 研究问题聚焦且有实际意义 ● ☒ 三种方法的组合有创新性 ● ☒ 理论文献支撑充分 ● ☒ 实验设计逻辑清晰 需要强化： ● ⚠ 伦理部分：需要更详细的隐私保护计划 ● ⚠ 时间线：需要具体日期和 contingency ● ⚠ 招募标准：需要可操作的筛选问卷 ● ⚠ 数据分析：需要明确统计方法和质性编码框架 请逐一回答上述问题，我会根据你的回复生成完整的研究计划底稿。特别是： 1) 参与者的具体筛选标准 2) 是否需要 REB 审批 3) AI 工具的最终选择 4) 具体的时间节点 准备好后请告诉我! 🎯	● Specific start and end dates (e.g., January Xth - January Xth) ● What is the deliverable for each milestone? ● Is there a contingency plan (what if recruitment is delayed?) 6. Prioritization of research objectives You mentioned multiple research questions (RQs). You need to clarify: ● What is the primary RQ? What are the secondary RQs? ● Which RQ does each of the three research methods primarily answer? Supplementary Confirmation Questions A. Completeness of Test Materials ● What exactly are the “Errors and Anomalies” in cheese_nightmare_data.csv? (Missing values? Outliers? Formatting errors?) ● Does the task instruction include success criteria? (e.g., “Correctly identified X outliers”) B. Semi-structured Interview Outline Can you list 3-5 core questions? For example: 1) Which step confused you the most? 2) How would you like the software improved? 3) … C. PURE Scoring Criteria ● What are the scoring scales for “Discoverability” and “Usability”? (1-5 points? 1-7 points?) ● How granular is the Standard Operating Procedure (SOP)? (How many steps are there in total?) D. Specific Implementation of Heuristic Evaluation ● Were all 10 heuristic principles used?
---	--

	• What is the definition of the severity score (0-4)? • Was the evaluation conducted independently or through group discussion? My Initial Suggestions Strengths to Maintain: •  Focused and practically relevant research question •  Innovative combination of three methods •  Strong theoretical and literature support •  Clear and logical experimental design Areas for Improvement: •  Ethics: A more detailed privacy protection plan is needed •  Timeline: Specific dates and contingencies are needed •  Recruitment Criteria: An operational screening questionnaire is needed •  Data Analysis: Clear statistical methods and qualitative coding framework are needed Please answer the above questions one by one. I will generate a complete research proposal draft based on your responses. Specifically: 1) Specific selection criteria for participants 2) Whether REB approval is required 3) Final selection of AI tools 4) Specific timelines Let me know when you are ready! 
用户测试组：STEM 专业背景，并且上过统计学课程，对统计概念有基本理解。招募渠道通过社交媒体，靠工作和专业背景进行筛选。因为是亲友测试所以筛选过程相对会简单一些。	User Testing Team: Candidates should have a STEM background, have taken statistics courses, and have a basic understanding of statistical concepts. Recruitment will be conducted through social media, with screening based on work experience and professional background.

UX「专家」的话我打算找同班同学，都是上了一学期课的同学，问题不大，我会从同班同学中筛选有 STEM 背景且大学修过统计课的人（比如心理学专业）。

数据全部都是匿名化的，记录表和视频会用匿名编号关联，所有数据会在研究执行完毕后全部删除。

模型方面，我决定使用 Gemini Computer Use，一个模型，执行三十次。

主要研究问题是 Jamovi 在进行数据分析过程中是否有可用性（我觉得这个问题可以被表述的更好）其余问题是次要问题。

根据文献，SUS 需要 70 分以上（Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale 这篇论文支持此观点），NASA-TLX 的阈值是 54 分（根据 HOW HIGH IS HIGH? A META-ANALYSIS OF NASA-TLX GLOBAL WORKLOAD SCORES 这篇论文当中的数据得出的结论）

访谈转录后的编码采用主题分析，通过 LLM 整理出来访谈当中提到的焦点问题，再根据这些焦点问题整理 Affinity Map 的核心主题，将用户的困惑和 quote 填进去。由 LLM 独立编码三十次应当能保证信度。

关于测试材料方面，会有一个极大离群值（输入错误）和一个格式错误（把数字打成字母）任务指导书只告诉对方要清理数据，不给成功标准。

访谈设计方面，遵循由宽泛到具体的顺序，大体方向行的问题包括：

总体而言，您如何评价刚才使用 Jamovi 的体验？[让受访者自由表达，若回答过简则进一步追问]“哪项任务最简单？为什么？”“哪项任务最困难或最令人沮丧？能否详细说明挑战性体现在哪里？”“是否有感到困惑或卡住的时刻？当时发生了什么？”“是否有对下一步操作感到不确定的时刻？”

Since this is a friend-and-family test, the screening process will be relatively simple.

For UX “experts,” I plan to recruit classmates who have taken the course for a semester. This shouldn’t be a problem. I will select classmates with STEM backgrounds and who have taken statistics courses in university (e.g., psychology majors).

All data will be anonymized. Record sheets and videos will be linked using anonymous IDs. All data will be deleted after the research is completed.

Regarding the model, I have decided to use Gemini Computer Use, one model, executed thirty times.

The primary research question is whether Jamovi has usability issues during data analysis (I think this question could be better phrased). Other questions are secondary.

According to the literature, a SUS score of 70 or higher is required (supported by the paper “Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale”), and the NASA-TLX threshold is 54 (based on data from the paper “HOW HIGH IS HIGH? A META-ANALYSIS OF NASA-TLX GLOBAL WORKLOAD SCORES”).

Post-transcription coding employed topic analysis. LLM was used to identify key questions raised during the interviews, and then an Affinity Map was created based on these key questions, incorporating the user’s confusion and quotes. Thirty independent coding attempts by LLM should ensure reliability.

Regarding the test materials, there will be one large outlier (input error) and one formatting error (numbers typed as letters). The task instructions only tell

PURE 上用五点评分, 颗粒度尽可能细, 比如, 能够正确找到 X 功能, 能够正确完成 X 操作。 十条启发式评估都会用, 专家是独立评估不是一起讨论 严重性的定义如下: 0 = 非可用性问题 1 = 表面问题 (若有时间可修复) 2 = 次要可用性问题 (低优先级修复) 3 = 重大可用性问题 (高优先级修复) 4 = 可用性灾难 (必须修复) 因为这个研究不涉及到医疗或者心理等敏感领域, 所以不会走 REB, 但是会严格遵循 TCPS 2 标准 (https://tcps2core.ca/) 最终的时间节点我希望你参考课程 Syllabus 帮我定一下, 上面有作业死线。	the interviewee to clean the data, without providing success criteria. In terms of interview design, a broad-to-specific approach was followed, with general questions including: Overall, how would you rate your experience using Jamovi? [Allow respondents to express themselves freely; if the answer is too brief, follow up with further questions.] “Which task was the easiest? Why?” “Which task was the most difficult or frustrating? Can you elaborate on where the challenge lay?” “Were there any moments when you felt confused or stuck? What happened then?” “Were there any moments when you felt uncertain about the next step?” Use a five-point rating system on PURE, with the granularity as fine as possible. For example, be able to correctly locate function X, be able to correctly complete operation X. All ten heuristics will be used; experts will evaluate independently, not in group discussions. Severity is defined as follows: 0 = Non-usability issue 1 = Superficial issue (can be fixed if time permits) 2 = Minor usability issue (low priority to fix) 3 = Major usability issue (high priority to fix) 4 = Usability catastrophe (must be fixed)
---	---

	Because this research does not involve sensitive areas such as medicine or psychology, it will not go through REB (Research by Entity) but will strictly follow the TCPS 2 standard (https://tcps2core.ca/). I hope you can refer to the course syllabus to help me set the final deadlines; there are assignment deadlines there.
--	--

Output Markdown:

Research Plan: Evaluating the Usability of Jamovi for One-Way ANOVA Analysis

****Course:**** UX604: Research Methods II
****Student:**** [Your Name]
****Submission Date:**** January 20, 2026

1. Research Overview

1.1 Problem Statement and Context

Statistical software interfaces, particularly those following the SPSS-inspired design paradigm, present significant usability challenges that can lead to methodological errors in research practice. Jamovi, an open-source statistical platform, exemplifies this design tradition while claiming to offer improved user experience. However, critical questions remain regarding whether its interface genuinely facilitates or impedes accurate statistical analysis.

One-way Analysis of Variance (ANOVA) represents an ideal test case for usability evaluation: it requires sufficient analytical complexity to expose interface weaknesses while remaining accessible to researchers across disciplines. The complete workflow—data import, cleaning, assumption checking, analysis execution, and interpretation—creates multiple points where poor interface design can induce user error.

This research addresses a fundamental gap in understanding how statistical software design impacts research quality. Given Jamovi's positioning as a free, accessible alternative to commercial platforms, understanding its usability has implications for research reproducibility, educational contexts, and open science initiatives.

1.2 Research Objectives

****Primary Research Question:****

To what extent does Jamovi's user interface facilitate or hinder the accurate completion of one-way ANOVA analysis workflows?

****Secondary Research Questions:****

1. What specific usability issues emerge at different stages of the statistical analysis workflow (data import, cleaning, analysis, visualization, interpretation)?
2. How do usability perceptions and problem discovery differ between general STEM users and UX-trained evaluators?
3. What are the capabilities and limitations of AI-assisted usability evaluation compared to human-centered methods?

1.3 Research Outputs

This progressive research plan will produce three deliverable reports:

1. ****Usability Study Report**** (Due: February 10, 2026)
 - Task performance metrics (completion time, error rates, success rates)
 - Standardized usability measurements (SUS, NASA-TLX)
 - Qualitative insights from think-aloud protocols and interviews
 - Thematic analysis of usability barriers
2. ****PURE Study Report**** (Due: March 10, 2026)
 - Expert heuristic evaluation findings with severity ratings
 - Comparative analysis of expert vs. novice problem discovery
 - Detailed step-by-step discoverability and ease-of-use ratings
 - Prioritized recommendations for interface improvements
3. ****AI Evaluation Study Report**** (Due: March 24, 2026)
 - Automated usability testing results across 30 independent trials
 - AI-generated heuristic evaluations and affinity mapping
 - Comparative synthesis across all three methodologies
 - Critical assessment of AI's role in UX research workflows

2. Research Methods

2.1 Study 1: User-Centered Usability Testing

2.1.1 Method Selection and Justification

This study employs ****mixed-methods usability testing**** with concurrent think-aloud protocol (Ericsson & Simon, 1993). This approach is justified by:

- ****Industry standard:**** User testing with real participants remains the gold standard for discovering genuine usability issues (Nielsen & Mack, 1994)
- ****Think-aloud validity:**** Concurrent verbalization effectively captures cognitive processes during task execution without significantly altering performance (Ericsson, 2006)
- ****Sample size rationale:**** Five participants can uncover approximately 85% of usability problems for interfaces of moderate complexity (Virzi, 1992; Nielsen & Landauer, 1993)
- ****Mixed-method value:**** Combining quantitative metrics with qualitative insights provides both statistical benchmarks and contextual understanding (Lazar et al., 2017)

2.1.2 Participants (n=5)

****Inclusion Criteria:****

- STEM disciplinary background (science, technology, engineering, mathematics)
- Completed at least one university-level statistics course
- Basic understanding of statistical concepts (hypothesis testing, p-values, variance)
- No formal UX training or professional UX experience
- Able to participate via Zoom with reliable internet connection

****Exclusion Criteria:****

- Current UX students or professionals
- Extensive prior Jamovi experience (>5 hours of use)
- Statistical software developers

****Recruitment Strategy:****

- Recruitment via social media networks and personal/professional contacts
- Screening based on educational background and professional experience
- Target diversity in STEM disciplines (mix of natural sciences, engineering, health sciences)

****Ethical Considerations:****

- Research adheres to TCPS 2 (Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans)
- Informed consent obtained prior to participation using adapted digital.gov template
- Full disclosure of recording, data usage, and anonymization procedures
- Participants may withdraw at any time without penalty
- No REB submission required (non-sensitive, non-medical research with adult volunteers)

2.1.3 Testing Protocol

****Pre-Test Phase (5 minutes):****

- Welcome and rapport building
- Review and sign informed consent form (electronic signature)
- Brief explanation of think-aloud protocol with demonstration
- Equipment check (Zoom recording, screen sharing, audio quality)

****Task Execution Phase (20-30 minutes):****

Participants assume the role of a sleep researcher investigating the question: "Does eating cheese before bed cause nightmares?" They complete four sequential tasks using Jamovi:

****Task 1: Data Import and Validation** (~5 min)**

Instruction: "Import the file cheese_nightmare_data.csv into Jamovi and verify the data loaded correctly."
Success criteria: File imported; participant confirms data structure

****Task 2: Data Cleaning** (~5-8 min)**

Instruction: "Examine the dataset for errors or anomalies. Clean the data as appropriate for analysis."
Hidden challenges: Dataset contains one extreme outlier (data entry error) and one format error (letter in numeric field)
Success criteria: Participant identifies and addresses data quality issues

****Task 3: One-Way ANOVA Execution** (~8-12 min)**

Instruction: "Conduct a one-way ANOVA comparing nightmare frequency across different cheese types. Provide statistical interpretation of your findings."
Success criteria: Correct ANOVA specification; appropriate interpretation of results

****Task 4: Data Visualization** (~5-8**

min)

Instruction: "Create an appropriate chart that clearly displays the relationship between cheese type and nightmare occurrence."
Success criteria: Suitable graph type selected; proper formatting for interpretation

****Post-Task Assessment (5 minutes):****

- System Usability Scale (SUS) - 10 items
- NASA Task Load Index (NASA-TLX) - 6 dimensions

****Semi-Structured Interview (10 minutes):****

Follow-up questions exploring user experience:

1. "Overall, how would you describe your experience using Jamovi for this analysis?"
2. "Which task was easiest? Why?"
3. "Which task was most difficult or frustrating? What specifically made it challenging?"
4. "Were there moments when you felt confused or stuck? What was happening at that point?"
5. "Were there times when you felt uncertain about what to do next? Can you describe those situations?"
6. [Probe specific observed behaviors:] "I noticed you [specific action]. Can you tell me what you were thinking at that moment?"

****Closing (2-3 minutes):****

- Thank participant for their time
- Provide compensation/acknowledgment
- Stop recording and save files

2.1.4 Data Collection Instruments

****Quantitative Metrics:****

1. ****Task Performance:****
 - Completion time per task (timestamp from instruction delivery to participant declaration of completion)
 - Task success rate (binary: complete/incomplete)
 - Error count (number of incorrect actions or backtracking instances)
2. ****Standardized Assessments:****
 - ****System Usability Scale (SUS):**** 10-item questionnaire; scores range 0-100; threshold for acceptable usability: ≥ 70 (Bangor et al., 2008)
 - ****NASA Task Load Index (NASA-TLX):**** 6-dimension workload assessment; threshold for high workload: >54 (Grier, 2015)

****Qualitative Data:****

1. ****Think-Aloud Transcripts:**** Audio recordings transcribed verbatim; analyzed using thematic analysis
2. ****Semi-Structured Interview Data:**** Coded for recurring themes, pain points, and mental models

3. ****Observer Notes:**** Structured observation following digital.gov guidelines, documenting:

- Critical incidents (errors, confusion, recovery attempts)
- Emotional responses (frustration, satisfaction, surprise)
- Navigation patterns and workflow strategies
- Unsolicited comments or questions

****Testing Environment:****

- Remote participation via Zoom (ecological validity through participant's own environment)
- Participants use personal computers (Windows/Mac compatibility check during recruitment)
- Jamovi pre-installed by participant with setup verification
- Screen recording and audio capture via Zoom cloud recording

****Test Materials Provided:****

- cheese_nightmare_data.csv (100 simulated observations: cheese_type [4 levels: Cheddar, Brie, Swiss, Blue], nightmare_frequency [continuous])
- Task instruction document (PDF)
- SUS questionnaire (Microsoft Forms)
- NASA-TLX questionnaire (Microsoft Forms)
- Informed consent form (PDF with electronic signature capability)

2.1.5 Data Analysis Plan

****Quantitative Analysis:****

- Descriptive statistics for completion times, error rates, SUS scores, NASA-TLX dimensions
- Individual task performance profiles
- Correlation analysis between subjective ratings and objective performance

****Qualitative Analysis:****

- ****Thematic Analysis**** of think-aloud and interview transcripts:
 1. Familiarization with data (read all transcripts)
 2. Initial code generation (line-by-line coding)
 3. Theme identification (group codes into candidate themes)
 4. Theme review and refinement
 5. Theme definition and naming
 6. Report writing with illustrative quotes

- ****AI-Assisted Coding Reliability:****

- Gemini 2.5 Pro Thinking processes all transcripts 30 times independently
- Compare thematic consistency across runs
- Final themes represent convergent findings across iterations

- ****Affinity Mapping:****

- Extract key user quotes and observations
- Group related issues into thematic clusters
- Identify patterns in pain

points, confusion, and success moments

- Create visual affinity diagram linking user needs to interface failures

2.2 Study 2: Pragmatic Usability Rating by Experts (PURE)

2.2.1 Method Selection and Justification

The PURE methodology bridges traditional expert evaluation and user testing by having UX professionals complete actual tasks while applying expert heuristics (Bligård & Osvalder, 2013). This approach is justified by:

- **Theory-Practice Integration:** Experts experience firsthand usability issues while applying theoretical frameworks
- **Problem Discovery Enhancement:** Trained evaluators can identify issues that novice users may normalize or fail to articulate (Nielsen, 1994)
- **Cost-Effectiveness:** 3-5 expert evaluators can discover 60-75% of usability issues with significantly lower resource requirements than extensive user testing (Nielsen & Molich, 1990; Jeffries et al., 1991)
- **Complementary Insights:** Expert perspectives reveal design pattern violations and accessibility issues beyond individual user experiences

2.2.2 Evaluators (n=3)

****Inclusion Criteria:****

- Current enrollment in UX Design graduate program (minimum one semester completed)
- STEM undergraduate background (psychology, cognitive science, or related fields preferred)
- Completed university-level statistics course
- Familiar with heuristic evaluation methodology
- No professional relationship with Jamovi development

****Exclusion Criteria:****

- Participants from Study 1 (user testing)
- Jamovi development team members or contributors
- Less than one semester of formal UX training

****Recruitment Strategy:****

- Recruit from UX604 cohort (Winter 2026)
- Target students with quantitative research backgrounds (psychology, HCI, data science)
- Ensure evaluator independence (no collaboration during assessment)

****Ethical Considerations:****

- Same informed consent and data protection protocols as Study 1
- TCPS 2 compliance maintained
- Screen recordings anonymized with participant codes

2.2.3 Testing Protocol

****Phase 1: Task Completion with Think-Aloud (20-30 minutes)****

Evaluators complete identical tasks as Study 1 participants:

- Task 1: Data Import and Validation
- Task 2: Data Cleaning
- Task 3: One-Way ANOVA Execution
- Task 4: Data Visualization

Concurrent verbalization captures expert mental models and immediate reactions.

****Phase 2: Step-by-Step PURE Evaluation (15-20 minutes)****

Evaluators review a standardized Standard Operating Procedure (SOP) document breaking the workflow into granular steps (approximately 25-30 steps total). For each step, evaluators rate:

****Discoverability**** (5-point scale):

- 5 = Immediately obvious, clear visual cues
- 4 = Easy to find with minimal exploration
- 3 = Requires moderate searching or menu navigation
- 2 = Hidden, requires significant effort to locate
- 1 = Nearly impossible to discover without external guidance

****Ease of Use**** (5-point scale):

- 5 = Intuitive, requires no prior knowledge
- 4 = Easy with minimal learning curve
- 3 = Manageable with some trial-and-error
- 2 = Difficult, frustrating, error-prone
- 1 = Extremely challenging, likely to cause errors

Example SOP steps:

- Step 3: "Locate the 'Import' function in Jamovi"
- Step 7: "Identify variables in the data panel"
- Step 12: "Access the ANOVA analysis menu"
- Step 18: "Specify the dependent variable"
- Step 23: "Interpret the ANOVA output table"

****Phase 3: Heuristic Evaluation (20-25 minutes)****

Independent evaluation against ****Nielsen's 10 Usability Heuristics:****

1. ****Visibility of System Status:**** System keeps users informed about actions through timely feedback
2. ****Match Between System and Real World:**** System uses familiar language and follows real-world conventions
3. ****User Control and Freedom:**** Users can easily undo/redo actions; clear emergency exits
4. ****Consistency and Standards:****

Platform conventions and terminology are consistent

5. ****Error Prevention:**** Design prevents problems or checks for them before commitment
6. ****Recognition Rather Than Recall:**** Objects, actions, options are visible; minimize memory load
7. ****Flexibility and Efficiency of Use:**** Accelerators for expert users; customization options
8. ****Aesthetic and Minimalist Design:**** Interface contains only relevant information
9. ****Help Users Recognize, Diagnose, and Recover from Errors:**** Error messages in plain language suggesting solutions
10. ****Help and Documentation:**** Easy to search, focused on user's task, concrete steps

****For each usability issue identified, evaluators document:****

- ****Description:**** Clear statement of the problem
- ****Location:**** Specific interface element or workflow stage
- ****Heuristic Violation(s):**** Which principle(s) are violated
- ****Severity Rating:****
 - 0 = Not a usability problem
 - 1 = Cosmetic issue (fix if time permits)
 - 2 = Minor usability problem (low priority)
 - 3 = Major usability problem (high priority)
 - 4 = Usability catastrophe (must fix before release)
- ****Recommended Fix:**** Specific, actionable improvement suggestion

****Phase 4: Post-Evaluation (5 minutes)****

- SUS questionnaire
- NASA-TLX assessment
- Brief reflection on evaluation process

2.2.4 Data Collection

****Quantitative Data:****

- Task performance metrics (identical to Study 1)
- PURE ratings: Discoverability and Ease of Use scores for each SOP step
- Heuristic violation counts by category
- Severity distribution of identified issues
- SUS and NASA-TLX scores

****Qualitative Data:****

- Think-aloud transcripts during task execution
- Detailed heuristic evaluation reports
- Recommended improvements with rationale
- Observer notes on expert strategies

2.2.5 Data Analysis Plan

****Quantitative Analysis:****

- Aggregate PURE ratings across evaluators (mean, SD, inter-rater reliability)
- Identify workflow steps with lowest discoverability/ease scores
- Frequency analysis of heuristic

violations - Severity-weighted problem prioritization - Comparative analysis: Expert vs. novice performance metrics **Qualitative Analysis:** - Thematic analysis of expert think-alouds (compare to Study 1 novice patterns) - Content analysis of heuristic evaluation reports - Categorization of recommended improvements by impact and feasibility - Synthesis of expert mental models and problem-solving strategies **Cross-Study Synthesis:** - Issues identified only by experts vs. only by users vs. both - Alignment between user-reported frustrations and heuristic violations - Expert prediction accuracy for novice struggles --- ### 2.3 Study 3: AI-Assisted Usability Evaluation #### 2.3.1 Method Selection and Justification This study explores emerging capabilities of AI agents in usability assessment through: **Automated Task Execution:** Gemini 2.0 Flash with computer use capabilities independently completes identical tasks 30 times **AI-Generated Heuristic Analysis:** Large language models evaluate interface against established usability principles **Computational Thematic Analysis:** LLMs synthesize qualitative data from Studies 1 and 2 **Rationale:** - **Scalability:** AI enables large-scale independent replications impossible with human participants, generating robust statistical distributions - **Consistency Control:** Each trial begins with cleared context, ensuring independence and comparability - **Emerging Practice:** As AI increasingly integrates into UX workflows, empirical assessment of its capabilities addresses both theoretical and practical knowledge gaps - **Researcher Goals Alignment:** Developing AI-native research tools benefits from understanding AI's evaluation strengths and limitations - **Methodological Innovation:** Triangulating human and AI insights may reveal blind spots in traditional methods #### 2.3.2 AI Tools and Configuration **Primary Tool: Gemini 2.0 Flash with Computer Use**	- Model: `gemini-2.0-flash-exp` - Access: Google AI Studio free tier API - Context reset: New session for each of 30 independent trials - Screen resolution: 1920x1080 (standardized) - Operating system: Ubuntu 22.04 LTS (virtual machine) **Execution Protocol:** **Phase 1: Automated Task Completion (30 Independent Trials)** For each trial, the AI agent receives identical instructions as human participants: <pre> ''' You are a sleep researcher analyzing the relationship between cheese consumption and nightmares. Using the statistical software Jamovi, complete the following tasks: 1. Import cheese_nightmare_data.csv and verify correct loading 2. Identify and clean any data errors or anomalies 3. Conduct a one-way ANOVA comparing nightmare frequency across cheese types 4. Create an appropriate visualization of the findings Document each step you take and any difficulties encountered. ''' </pre> **Data Collected Per Trial:** - Task completion status (success/failure for each of 4 tasks) - Step-by-step action log (mouse clicks, keyboard inputs, menu navigation) - Time elapsed per task - Errors and recovery attempts - Final ANOVA results and interpretation accuracy **Phase 2: AI-Generated Heuristic Evaluation (30 Independent Runs)** Following task completion, Gemini analyzes Jamovi interface screenshots and workflow documentation using this prompt: <pre> ''' You are a UX expert evaluator. Analyze the Jamovi statistical software interface based on your experience completing the tasks. Evaluate against Nielsen's 10 Usability Heuristics: 1. Visibility of system status 2. Match between system and real world 3. User control and freedom 4. Consistency and standards 5. Error prevention 6. Recognition rather than recall 7. Flexibility and efficiency of use 8. Aesthetic and minimalist design 9. Help users recognize, diagnose, recover from errors 10. Help and documentation For each usability issue identified: </pre>	- Describe the specific problem - Identify which heuristic(s) it violates - Rate severity (0-4 scale): 0 = Not a problem 1 = Cosmetic issue 2 = Minor problem 3 = Major problem 4 = Catastrophic problem - Suggest specific improvement Provide structured output in JSON format. <pre> ''' </pre> **Phase 3: Computational Thematic Analysis** Gemini processes transcripts from Studies 1 and 2 to: **Generate Affinity Map (30 Independent Runs):** <pre> ''' Analyze the following usability test transcripts. Identify: - Common pain points mentioned by multiple participants - Recurring themes in user frustration or confusion - Patterns in successful vs. unsuccessful strategies - Key quotes illustrating major usability issues Group related findings into thematic clusters. For each cluster: - Theme name - Supporting quotes (with participant IDs) - Frequency count - Relationship to task stages Output as structured affinity map in JSON. ''' </pre> **Cross-Method Synthesis:** <pre> ''' Compare findings from: - Novice user testing (n=5) - Expert PURE evaluation (n=3) - AI task execution (n=30) Identify: - Issues discovered by all three methods - Issues unique to each method - Discrepancies in severity assessment - Complementary insights across approaches Synthesize into unified usability problem set with evidence source tagging. ''' </pre> #### 2.3.3 Data Analysis Plan **Automated Task Performance:** - Success rate distribution across 30 trials (binomial test for each task) - Completion time descriptive statistics - Error pattern frequency analysis - Comparison to human performance (Studies 1 & 2)
--	---	--

**AI Heuristic Evaluation

Synthesis:**

- Aggregate heuristic violations across 30 runs (mode, frequency)
- Severity rating consistency (inter-run agreement)
- Compare AI-identified issues to human expert findings
- Assess AI's ability to match expert judgment

Thematic Analysis Validation:

- Convergence analysis: theme stability across 30 independent LLM runs
- Compare AI-generated themes to researcher's manual coding
- Identify novel insights vs. confirmation of human findings
- Evaluate AI's capacity for nuanced qualitative interpretation

Methodological Comparison:

Dimension	User Testing	PURE	AI Evaluation
Problem discovery rate	100%	95%	90%
% of total issues	100%	95%	90%
% of total issues	100%	95%	90%
False positive rate	5%	10%	15%
Issues not confirmed	5%	10%	15%
Issues not confirmed	5%	10%	15%
Severity calibration	High	Medium	Low
Alignment with expert consensus	High	Medium	Low
Alignment with user impact	High	Medium	Low
Alignment with both	High	Medium	Low
Cost efficiency	High	Medium	Low
Time x participants	High	Medium	Low
Time x evaluators	High	Medium	Low
Computational cost	High	Medium	Low
Scalability	Limited	Limited	Limited
Limited by recruitment	Limited	Limited	Limited
Limited by expertise	Limited	Limited	Limited
Essentially unlimited	Limited	Limited	Limited
Contextual insight	Rich	Medium	Low
Rich narratives	Rich	Medium	Low
Theory-grounded	Rich	Medium	Low
Pattern recognition	Rich	Medium	Low

Deliverable Outputs:

1. Quantitative performance comparison matrix
2. Unified usability problem taxonomy with multi-method evidence
3. Critical assessment of AI's role in UX research
4. Practical guidelines for hybrid human-AI evaluation workflows

3. Research Timeline

3.1 Detailed Schedule

Based on course deadlines (from syllabus):

- Assignment #1 (Research Plan): Due Tuesday, January 20, 2026
- Assignment #2 (Usability Study): Due Tuesday, February 10, 2026
- Assignment #3 (PURE Study): Due Tuesday, March 10, 2026
- Assignment #4 (AI Evaluation): Due Tuesday, March 24, 2026

Phase	Dates	Deliverables	Time Allocation
Research Plan Development	Jan 6-20, 2026	Finalize methodology	1 week

- Prepare test materials
- Develop consent forms & questionnaires
- Submit

Research Plan (Jan 20)** | 2 weeks |

****Study 1: User Testing**** | Jan 21-Feb 10, 2026 | | |
 | 4 Participant recruitment | Jan 21-27 | • Screen 8-10 candidates
 • Confirm 5 participants
 • Schedule sessions | 1 week |
 | 4 Data collection | Jan 28-Feb 3 |
 • Conduct 5 usability sessions (1/day)
 • Observer notes & recordings | 1 week |
 | 4 Analysis & reporting | Feb 4-10 |
 • Transcribe sessions
 • Thematic analysis
 • Calculate metrics
****Submit Usability Study Report (Feb 10)**** | 1 week |
****Reading Week**** | Feb 11-17 | (No research activities) | 1 week |
****Study 2: PURE Evaluation**** | Feb 18-Mar 10, 2026 | | |
 | 4 Evaluator recruitment | Feb 18-23 | • Recruit 3 UX students
 • Verify STEM backgrounds
 • Schedule sessions | <1 week |
 | 4 Data collection | Feb 24-Mar 2 |
 • Conduct 3 expert evaluations (1/day)
 • Heuristic analysis sessions | 1 week |
 | 4 Analysis & reporting | Mar 3-10 |
 • Aggregate PURE ratings
 • Synthesize heuristic findings
 • Comparative analysis with Study 1
****Submit PURE Study Report (Mar 10)**** | 1 week |
****Study 3: AI Evaluation**** | Mar 11-24, 2026 | | |
 | 4 AI infrastructure setup | Mar 11-12 | • Configure Gemini API
 • Test virtual environment
 • Validate prompts | 2 days |
 | 4 Automated data collection | Mar 13-16 | • Run 30 AI task trials
 • Run 30 heuristic evaluations
 • Run 30 thematic analysis iterations | 4 days |
 | 4 Analysis & synthesis | Mar 17-24 |
 • Statistical analysis of AI performance
 • Validate AI-generated themes
 • Cross-method comparison
 • Final synthesis & recommendations
****Submit AI Evaluation Report (Mar 24)**** | 1 week |

3.2 Contingency Planning

Recruitment Challenges:

- ***Risk:** Insufficient user testing participants by Jan 27
- ***Mitigation:** Begin recruitment immediately upon research plan approval; maintain pool of 8-10 screened candidates; offer flexible scheduling

Technical Issues:

- ***Risk:** Jamovi software bugs or Zoom recording failures
- ***Mitigation:** Conduct pilot test before first participant; have backup recording method (OBS Studio); troubleshoot guide for common issues

Data Quality Concerns:

- ***Risk:** Participants unfamiliar with think-aloud protocol provide limited verbalization
- ***Mitigation:** Provide clear demonstration and practice task; gentle prompting during sessions

("What are you thinking right now?")

AI Limitations:

- ***Risk:** Gemini Computer Use fails to complete tasks or API quota exceeded
- ***Mitigation:** Budget analysis based on free tier limits (15 requests/minute); distribute trials over multiple days; fallback to manual task simulation if necessary

Analysis Timeline Pressure:

- ***Risk:** Transcription and coding delays
- ***Mitigation:** Begin analysis immediately after each session rather than batching; use AI-assisted transcription (Gemini Audio); prioritize critical findings if time-constrained

3.3 Milestones and Checkpoints

- ****Jan 20:**** Research plan approved → green light to begin recruitment
- ****Jan 27:**** All Study 1 participants confirmed → proceed with testing
- ****Feb 3:**** Study 1 data collection complete → begin transcription
- ****Feb 10:**** Study 1 report submitted → assess findings to refine Study 2
- ****Feb 23:**** Study 2 evaluators confirmed → proceed with expert sessions
- ****Mar 2:**** Study 2 data collection complete → begin synthesis
- ****Mar 10:**** Study 2 report submitted → finalize AI prompts based on insights
- ****Mar 16:**** AI data collection complete → begin validation analysis
- ****Mar 24:**** Final report submitted → research complete

4. Participants, Evaluators, and Tools

4.1 Study 1: User Testing Participants (n=5)

4.1.1 Target Population

Demographic Profile:

- STEM professionals or graduate students
- Age range: 22-45 years (typical range for early to mid-career researchers)
- Mix of genders (target: minimum 40% underrepresented gender)
- Diversity of STEM disciplines (target: no more than 2 from same field)

Prerequisite Knowledge:

- Completed university-level statistics course (introductory or advanced)
- Familiar with core concepts: hypothesis testing, p-values, ANOVA principles
- Basic computer literacy (file management, software installation)
- No professional UX training or experience

Jamovi Experience:

- None to minimal (<5 hours total usage)
- May have experience with other statistical software (SPSS, R, Excel)

4.1.2 Recruitment Process

Outreach Channels:

- Personal professional networks (colleagues, former classmates)
- University alumni groups (STEM departments)
- Social media (LinkedIn, Twitter/X with relevant hashtags: #DataScience, #Statistics)
- Snowball sampling (referrals from initial participants)

Screening Questionnaire:

1. What is your highest level of education completed?
 - ☐ Bachelor's degree
 - ☐ Master's degree
 - ☐ PhD
 - ☐ Other: _____
2. What is your field of study/work? (open text)
3. Have you completed a university-level statistics course?
 - ☐ Yes → Which course(s)? _____
 - ☐ No → [Exclude]
4. Rate your familiarity with these statistical concepts (1=Never heard of it, 5=Very comfortable):
 - Hypothesis testing: 1 2 3 4 5
 - P-values: 1 2 3 4 5
 - Analysis of Variance (ANOVA): 1 2 3 4 5

[Minimum threshold: Average ≥3]
5. Do you have professional experience in UX design or research?
 - ☐ Yes → [Exclude]
 - ☐ No
6. How many hours have you used Jamovi statistical software?
 - ☐ 0 hours
 - ☐ 1-5 hours
 - ☐ 6-10 hours → [Exclude]
 - ☐ 10+ hours → [Exclude]
7. Are you available for a 1-hour Zoom session during [date range]?
 - ☐ Yes → Preferred times: _____
 - ☐ No
8. Do you have access to:
 - ☐ Computer (Windows/Mac) with admin rights to install software
 - ☐ Stable internet connection for video call
 - ☐ Quiet space for 1 hour

Selection Criteria:

- First 5 qualified respondents meeting all criteria
- Backup pool of 3 additional qualified candidates in case of cancellations

4.1.3 Participant Tracking

Participant ID System:

- Anonymous codes: P01, P02, P03, P04, P05
- Master list (kept separate from data): links ID to contact information
- All data files, transcripts, recordings labeled with ID only

Tracking Spreadsheet Columns:

ID	Recruitment Date	Screening Score	Discipline	Session Date	Session Time	Consent Signed	Recording Saved	Transcription Complete	Compensation Sent
P01	Jan 22	Qualified	Biology	Jan 29, 10am	58 min	✓	✓	✓	✓

4.1.4 Informed Consent

Consent Form Elements (adapted from digital.gov template):

- ****Study Purpose:**** Evaluating the usability of Jamovi statistical software
 - ****Procedures:**** 1-hour Zoom session involving task completion, questionnaires, and interview
 - ****Recording:**** Video and audio will be recorded via Zoom; screen sharing will capture interface interactions
 - ****Risks:**** Minimal; potential for minor frustration during challenging tasks
 - ****Benefits:**** Contributing to open-source software improvement; learning Jamovi basics
 - ****Confidentiality:****
 - All data anonymized using participant IDs
 - Recordings stored on encrypted, password-protected devices
 - Only research team has access to identifiable data
 - Data retention: deleted within 2 weeks of course completion (April 15, 2026)
 - ****Voluntary Participation:****
 - Right to withdraw at any time without penalty
 - Right to skip questions or tasks
 - No impact on relationship with university
 - ****Compensation:**** Acknowledgment in research portfolio (no monetary compensation for friends/family testing)
 - ****Contact Information:**** Researcher contact for questions; TCPS 2 research ethics resources
 - ****Signature:**** Electronic signature with date constitutes informed consent
- **Consent Procedure:****
1. Form emailed 48 hours before session for review
 2. Reviewed verbally at session start; opportunity for questions
 3. Electronic signature collected via DocuSign/Adobe Sign
 4. Copy provided to participant; original retained by researcher
-

4.2 Study 2: PURE Evaluators (n=3)

4.2.1 Target Population

Qualification Requirements:

- Current enrollment in UX604 (Winter 2026) or equivalent UX graduate program
- Minimum one semester of formal UX training completed
- Undergraduate degree in STEM field (psychology, cognitive science, HCI, or related)
- Completed university-level statistics course
- Familiar with Nielsen's heuristic evaluation methodology
- No prior participation in Study 1

Desired Expertise:

- Understanding of human-computer interaction principles
- Experience conducting heuristic evaluations (coursework or professional)
- Ability to articulate design rationale and violations

4.2.2 Recruitment Process

Recruitment Approach:

- Direct recruitment from UX604 cohort (email to classmates)
- Brief presentation during class session explaining study purpose
- Screening via short questionnaire

Screening Questions:

1. Have you completed at least one semester of UX coursework?
 - ☐ Yes → Which courses? _____
 - ☐ No → [Exclude]
2. What was your undergraduate major? (open text)
 - [Verify STEM background]
3. Have you completed a university statistics course?
 - ☐ Yes → Course name/number: _____
 - ☐ No → [Exclude]
4. Are you familiar with Nielsen's 10 Usability Heuristics?
 - ☐ Very familiar (can name several)
 - ☐ Somewhat familiar (studied in class)
 - ☐ Not familiar → [Exclude]
5. Have you conducted a heuristic evaluation before?
 - ☐ Yes, multiple times (professional/academic)
 - ☐ Yes, once or twice (coursework)
 - ☐ No, but understand the methodology
 - ☐ No → [Provide brief training materials]
6. Did you participate in the user testing phase of this study (Study 1)?
 - ☐ Yes → [Exclude]
 - ☐ No
7. Availability for 90-minute Zoom session during [date range]?
 - ☐ Yes → Preferred times: _____

- [] No **Selection:** - First 3 qualified classmates meeting all criteria - Backup: 1-2 additional qualified candidates ### 4.2.3 Evaluator Tracking **Evaluator ID System:** - Anonymous codes: E01, E02, E03 - Separate master list linking IDs to identities - All evaluation reports, recordings labeled with ID only **Tracking Spreadsheet:**<table><tr><td>ID</td><td>UX Background</td><td>STEM Field</td><td>Session Date</td><td>PURE Ratings Complete</td><td>Heuristic Eval Complete</td><td>SUS/TLX Complete</td><td>Consent Signed</td></tr><tr><td>E01</td><td>UX604 + UX603</td><td>Psychology</td><td>Feb 25, 2pm</td><td>✓</td><td>✓</td><td>✓</td><td>✓</td></tr></table> ### 4.2.4 Informed Consent Same consent protocol as Study 1, with additional clarification: - Role as expert evaluator explained - Emphasis on independent assessment (no collaboration with other evaluators) - Heuristic evaluation reports may be quoted anonymously in research findings --- ### 4.3 Study 3: AI Tools Configuration ### 4.3.1 Primary AI Agent: Gemini 2.0 Flash with Computer Use **Technical Specifications:** - **Model:** `gemini-2.0-flash-exp` - **API Access:** Google AI Studio (free tier) - Rate limits: 15 requests per minute, 1500 requests per day - Context window: 1 million tokens - Computer use capability: Enabled - **Execution Environment:** - Virtual machine: Ubuntu 22.04 LTS - Screen resolution: 1920x1080 (standardized) - Jamovi version: Latest stable release (pre-installed) - Browser: Chrome (for potential web-based Jamovi access) **Session Management:** - Each of 30 trials uses fresh session (cleared context) - Systematic prompt versioning for reproducibility - Automated logging of all agent actions ### 4.3.2 AI Evaluation Prompts **Task Execution Prompt Template:**<pre>''' You are a sleep researcher investigating cheese consumption and nightmares. You have access to a computer with Jamovi statistical software installed.</pre>	ID	UX Background	STEM Field	Session Date	PURE Ratings Complete	Heuristic Eval Complete	SUS/TLX Complete	Consent Signed	E01	UX604 + UX603	Psychology	Feb 25, 2pm	✓	✓	✓	✓	Your research question: "Does eating cheese before bed cause nightmares?" Complete these tasks using Jamovi: TASK 1: Import and Validate Data - Open the file "cheese_nightmare_data.csv" - Verify the data loaded correctly (100 rows, 2 columns) - Confirm variable types are appropriate TASK 2: Clean the Data - Examine the dataset for errors, outliers, or anomalies - Identify and address any data quality issues you find - Document what you cleaned and why TASK 3: Conduct One-Way ANOVA - Compare nightmare frequency across different cheese types - Run appropriate statistical test - Interpret results (statistical significance, effect size) - State your conclusion about the research question TASK 4: Visualize Results - Create a chart that clearly shows the relationship between cheese type and nightmares - Ensure the visualization is properly labeled and interpretable Document each step you take. If you encounter difficulties, describe the issue and how you attempted to resolve it. <pre>'''</pre> **Heuristic Evaluation Prompt Template:**<pre>''' You are an expert UX evaluator. You have just completed four tasks using Jamovi statistical software. Based on this experience, evaluate the interface against Nielsen's 10 Usability Heuristics.</pre> CONTEXT: - Software: Jamovi (open-source statistical platform) - Tasks completed: Data import, cleaning, ANOVA analysis, visualization - User role: Researcher with statistics knowledge HEURISTICS TO EVALUATE: 1. Visibility of system status 2. Match between system and the real world 3. User control and freedom 4. Consistency and standards 5. Error prevention 6. Recognition rather than recall 7. Flexibility and efficiency of use 8. Aesthetic and minimalist design 9. Help users recognize, diagnose, and recover from errors 10. Help and documentation FOR EACH USABILITY ISSUE YOU IDENTIFY:<pre>{ "issue id": "unique identifier",</pre>	"description": "Clear description of the problem", "location": "Specific interface element or workflow stage", "heuristics_violated": ["list of violated heuristics"], "severity": 0-4, "severity_justification": "Why this severity rating", "user_impact": "How this affects user experience", "recommended_fix": "Specific, actionable improvement" } SEVERITY SCALE: 0 = Not a usability problem 1 = Cosmetic issue (fix if time permits) 2 = Minor usability problem (low priority) 3 = Major usability problem (high priority) 4 = Usability catastrophe (must fix) Provide your evaluation as a JSON array of issues. Focus on problems that genuinely impacted your ability to complete tasks efficiently and accurately. <pre>'''</pre> **Thematic Analysis Prompt Template:**<pre>''' You are a qualitative research analyst. Analyze the following usability test transcripts to identify recurring themes, pain points, and patterns.</pre> TRANSCRIPTS: [Insert all participant think-aloud and interview transcripts] YOUR TASK: 1. Identify common themes across participants related to: - Confusion or difficulty points - Frustration or dissatisfaction - Moments of success or ease - Unmet expectations - Suggestions for improvement 2. For each theme: - Provide a clear theme name and description - List supporting quotes with participant IDs (e.g., P02, E01) - Count how many participants mentioned this theme - Note which task stages the theme relates to - Assess severity/importance 3. Create an affinity map grouping related themes into clusters OUTPUT FORMAT (JSON):<pre>{ "themes": [{ "theme_name": "string", "description": "string", "task_stages": ["list"], "frequency": number, "severity": "high/medium/low", "supporting_quotes": [{"participant_id": "string", "quote": "string"}] }] }</pre>
ID	UX Background	STEM Field	Session Date	PURE Ratings Complete	Heuristic Eval Complete	SUS/TLX Complete	Consent Signed											
E01	UX604 + UX603	Psychology	Feb 25, 2pm	✓	✓	✓	✓											

```
}
],
"theme_clusters": [
{
"cluster_name": "string",
"themes": ["list of theme names in
this cluster"]
}
]
}
```

Focus on themes that appear across multiple participants and that have clear implications for interface redesign.

```
\\`\\`
```

4.3.3 Data Collection and Storage

****Automated Logging:****

- Each AI trial generates:
 - Timestamped action log (JSON)
 - Screenshot captures at key decision points
 - Success/failure status for each task
 - Total execution time
 - Error recovery attempts
 - Final analysis results (for accuracy verification)

****Storage Structure:****

```
\\`\\`
/ai_evaluation_data/
/trials/
/trial_001/
- actions.json
- screenshots/
- completion_status.json
- timing_log.json
/trial_002/
...
/trial_030/
/heuristic_evaluations/
/eval_001.json
...
/eval_030.json
/thematic_analyses/
/analysis_001.json
...
/analysis_030.json
\\`\\`
```

****Quality Control:****

- Automated validation of JSON outputs (schema checking)
- Manual review of first 3 trials to verify prompt effectiveness
- Mid-point check (trial 15) to assess consistency
- Final audit of all 30 trials for anomalies

4.3.4 Ethical Considerations for AI Use

****Transparency:****

- Full disclosure that AI tools are used for evaluation
- Clear documentation of prompts, model versions, and parameters
- Acknowledgment of AI's role in analysis sections of reports

****Bias Awareness:****

- AI reflects training data biases; findings interpreted with this caveat
- Cross-validation with human-generated analyses required
- AI used to augment, not replace,

human judgment

****Data Privacy:****

- No personally identifiable information from human participants included in AI prompts
- Transcripts anonymized before AI processing
- API usage complies with Google's terms of service

****Academic Integrity:****

- AI-generated content clearly attributed
- Critical analysis of AI outputs (not blindly accepted)
- AI serves as research tool, not substitute for researcher thinking

5. Budget and Resource Allocation

5.1 Financial Budget

****Total Estimated Cost: \$0 CAD****

Category	Item	Cost	Justification
Software	Jamovi	\$0	Open-source, freely available
	Zoom	\$0	Using existing university license
	Google Gemini API	\$0	Free tier sufficient for 30 trials
	Microsoft Forms	\$0	Included in university account
	Transcription	\$0	Using Gemini Audio (free tier)
	Participant Compensation		
	User testing (n=5)	\$0	Friends/family volunteer basis
	PURE evaluators (n=3)	\$0	Classmate exchange; reciprocal participation
Materials	Test dataset	\$0	Researcher-generated simulated data
	Consent forms	\$0	Digital templates (digital.gov)
	Questionnaires	\$0	Standard validated scales (public domain)
	Equipment		
	Researcher computer	\$0	Personal device
	Participants' computers	\$0	Bring your own device (BYOD)
Total		**\$0**	

5.2 Time Budget

****Total Researcher Time: ~100 hours****

Activity	Hours	Breakdown
Study 1: User Testing 35 hrs		
4 Preparation 8 hrs Materials creation, pilot test, recruitment		
4 Data collection 10 hrs 5 sessions × 2 hrs (session + setup)		
4 Transcription 5 hrs AI-assisted, requires verification		
4 Analysis 8 hrs Thematic coding, metrics calculation		
4 Report writing 4 hrs		
Study 2: PURE 30 hrs		
4 Preparation 5 hrs SOP document, evaluator briefing		
4 Data collection 9 hrs 3 sessions × 3 hrs (longer protocol)		

4 Analysis	10 hrs	Aggregate ratings, synthesize heuristics
4 Report writing	6 hrs	Comparative analysis with Study 1
Study 3: AI Evaluation 25 hrs		
4 Setup	6 hrs	VM configuration, prompt engineering
4 Monitoring trials	4 hrs	Supervision of automated runs
4 Analysis	10 hrs	Statistical analysis, validation
4 Report writing	5 hrs	
Cross-Study Synthesis 10 hrs		
Final integration, recommendations		
Total **100 hrs** ~8-9 hrs/week over 12-week term		

5.3 Resource Requirements

****Physical Resources:****

- Researcher laptop with webcam and microphone (owned)
- Reliable internet connection (university/home)
- External hard drive for backup storage (owned)
- Headphones for transcript review (owned)

****Digital Resources:****

- Zoom account with cloud recording (university license)
- Google Workspace (university account)
- Gemini AI Studio access (free Google account)
- Virtual machine or cloud compute instance (Google Colab free tier)
- Reference management software: Zotero (free)

****Human Resources:****

- 5 user testing participants (recruited from network)
- 3 UX evaluator participants (classmates)
- Course instructor (for research plan feedback and approval)
- No additional research assistants required

****Institutional Resources:****

- University library access for literature review
- MyLearningSpace for assignment submission
- Email for participant communication

5.4 Risk Mitigation and Contingencies

****Low-Cost Backup Plans:****

- **If Gemini API limits exceeded:****
 - Spread trials over multiple days
 - Create secondary Google account for additional free tier
 - Reduce from 30 to 20 trials if necessary (still statistically valid)
- **If participant recruitment falls short:****
 - Extend recruitment period by 3-5 days
 - Broaden STEM criteria to include adjacent fields (education

in STEM, science communication) - Leverage snowball sampling (ask participants for referrals) 3. **If Zoom recording fails:** - Use OBS Studio (free, open-source) as backup recorder - Participant's local recording as secondary backup - Detailed observer notes can partially compensate 4. **If transcription quality is poor:** - Manual transcription of critical sections - Focus on higher-quality audio segments - Supplementary observer notes capture key quotes --- ## 6. Data Management and Ethical Compliance ### 6.1 TCPS 2 Compliance This research adheres to the **Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (TCPS 2)**, specifically: **Core Principles:** **Respect for Persons:** Voluntary informed consent; right to withdraw **Concern for Welfare:** Minimal risk; protecting participant wellbeing **Justice:** Fair participant selection; equitable burden/benefit distribution **Rationale for No REB Submission:** - Research involves minimal risk (software usability testing) - No vulnerable populations targeted - No collection of sensitive personal information - No deception or psychological manipulation - Educational context (course assignment) - Qualifies for delegated review / exemption under TCPS 2 Article 2.5 **However, research still follows REB-level standards:** - Full informed consent process - Data protection and anonymization protocols - Right to withdraw without consequences - Secure data storage and destruction plan ### 6.2 Data Protection Plan **Anonymization:** - Participant IDs assigned upon consent (P01-P05, E01-E03) - Master list linking IDs to identities kept in separate encrypted file - All data files, transcripts, reports use IDs only - Demographic information reported in aggregate only (no individual identification possible) **Storage Security:**	- Recordings stored on password-protected, encrypted external hard drive - Cloud storage (Google Drive) uses university account with 2-factor authentication - No data stored on public or shared computers - Backup copy on second encrypted device **Access Control:** - Only researcher has access to identifiable data - Course instructor may review anonymized data for grading - No third-party access without explicit participant consent **Data Retention and Destruction:** - Active use: January 20 - March 24, 2026 (course duration) - Retention: Until April 15, 2026 (2 weeks post-course for grading) - Destruction: Permanent deletion of all recordings, transcripts, and identifiable information by April 15, 2026 - Anonymized summary data may be retained for portfolio use (no identifiers) ### 6.3 Participant Rights **Voluntary Participation:** - Participation entirely voluntary - No penalty for declining or withdrawing - No impact on relationship with university or researcher **Right to Withdraw:** - Participants may withdraw at any time before, during, or up to 1 week after their session - Withdrawal results in complete data deletion - No explanation required **Right to Skip:** - Participants may decline to answer any question - Participants may decline to complete any task - Partial participation acceptable **Debriefing:** - Participants receive summary of research findings upon request (after March 24, 2026) - Contact information provided for questions or concerns - Resources for participant support if any distress experienced (unlikely but available) --- ## 7. Expected Outcomes and Contributions ### 7.1 Anticipated Findings **Study 1 (User Testing):** - Identification of 10-15 distinct usability issues across Jamovi workflow - Average SUS score prediction: 60-75 (acceptable to good usability range) - Average NASA-TLX score prediction: 45-60 (moderate workload)	- Common pain points likely in: data type detection, ANOVA menu navigation, output interpretation **Study 2 (PURE):** - Expert discovery of 15-25 usability issues (broader than user testing due to systematic evaluation) - Lower PURE ratings expected for: discoverability of advanced options, consistency across modules - Heuristic violations predicted primarily in: #6 (Recognition over recall), #5 (Error prevention), #10 (Help/documentation) - Experts likely identify issues users normalized or couldn't articulate **Study 3 (AI Evaluation):** - AI task completion rate prediction: 60-80% (lower than humans due to interface complexity) - AI heuristic evaluation convergence expected after ~15-20 trials - Novel insight: AI may identify workflow inefficiencies humans adapt to unconsciously - Limitations: AI may miss context-dependent frustrations or emotional responses **Cross-Study Synthesis:** - Core issues identified by all three methods: ~5-8 critical problems - Method-unique findings: User testing (emotional impact), PURE (theory violations), AI (efficiency patterns) - Severity calibration differences: Users rate by frustration, experts by principle violation, AI by task failure rate ### 7.2 Practical Contributions **To Open-Source Community:** - Detailed usability report delivered to Jamovi development team - Prioritized improvement recommendations with evidence from 38 evaluation sessions - Potential GitHub issue submissions with user quotes and expert analysis **To UX Research Practice:** - Empirical comparison of three evaluation methods (cost, coverage, depth) - Guidelines for when to use each method or combine approaches - Documentation of AI's strengths/limitations in usability evaluation (emerging practice area) **To Course Portfolio:** - Comprehensive case study demonstrating: - Quantitative and qualitative UX research skills - Experiment design and execution - Multi-method triangulation - Critical evaluation of research methodologies - Professional reporting and visualization ### 7.3 Theoretical Contributions
---	--	--

****To UX Knowledge:****

- Insight into statistical software usability (underexplored domain compared to consumer apps)
- Understanding of domain-specific usability (statistical reasoning + interface interaction)
- Evidence on whether SPSS-inspired design paradigm serves modern users

****To Research Methodology:****

- Validation (or refutation) of AI as viable usability evaluation tool
- Understanding AI's complementary role vs. replacement potential
- Contribution to discourse on hybrid human-AI research workflows

8. Limitations and Constraints

8.1 Methodological Limitations

****Sample Size:****

- 5 user participants limits generalizability
- 3 expert evaluators may not capture full range of heuristic violations
- Findings indicative rather than definitive

****Participant Pool:****

- Convenience sampling (friends, classmates) introduces selection bias
- STEM background requirement excludes social science/humanities researchers who also use statistics
- Self-selection bias: volunteers may be more tech-savvy or interested in usability

****Task Scope:****

- Single statistical test (one-way ANOVA) doesn't represent full Jamovi functionality
- Simulated scenario may not capture real research pressures and stakes
- 1-hour session limits exposure to interface; doesn't assess learning curve over time

****Remote Testing:****

- Less control over testing environment (distractions, hardware variability)
- Reduced ability to observe subtle non-verbal cues
- Technical issues may confound usability issues

8.2 Resource Constraints

****Time:****

- 12-week semester compresses timeline; limited iteration possible
- Single researcher limits analysis depth (no inter-rater reliability for qualitative coding)
- Concurrent course demands may affect quality of analysis

****Budget:****

- Zero budget precludes participant incentives (may affect recruitment or motivation)
- No funding for professional transcription services or specialized analysis software
- Reliance on free-tier AI limits

number of trials and query complexity

****Expertise:****

- Researcher is graduate student, not seasoned UX professional
- First major usability study; learning curve may affect protocol execution
- Limited access to advanced statistical analysis tools or mentorship

8.3 Scope Boundaries

****What This Study Does NOT Cover:****

- Other Jamovi functions (regression, factor analysis, data visualization beyond basic charts)
- Longitudinal usability (learning curve over weeks/months)
- Accessibility for users with disabilities
- Cross-cultural usability (non-English interfaces, different statistical conventions)
- Expert statistician users (advanced use cases)
- Comparison to other statistical software (SPSS, R, SAS)

****Acknowledged Biases:****

- Researcher has prior Jamovi experience; may unconsciously guide participants
- Focus on errors may emphasize negative over positive aspects of interface
- Academic context may not reflect industry use pressures

9. Dissemination Plan

9.1 Academic Deliverables

****Required Course Submissions:****

1. Research Plan (This document) - January 20, 2026
2. Usability Study Report - February 10, 2026
3. PURE Study Report - March 10, 2026
4. AI Evaluation Study Report - March 24, 2026

****Optional Extensions:****

- Conference poster submission: UX research conferences (e.g., CHI Student Research Competition, local HCI symposium)
- Blog post: Medium or personal website sharing key findings and methodology
- Open-source contribution: Share anonymized data and analysis scripts on GitHub

9.2 Stakeholder Communication

****Jamovi Development Team:****

- Email summary of findings with link to full report (post-course, April 2026)
- GitHub issue submissions for critical usability bugs
- Offer to present findings via video call if interest exists

****Course Instructor:****

- Regular progress updates during

office hours

- Draft sections for feedback before final submissions
- Presentation of final synthesis (if opportunity exists)

****Participants:****

- Thank-you email with overview of contributions (post-study)
- Offer to share final anonymized findings (post-course)
- Acknowledgment in portfolio materials (with consent)

9.3 Portfolio Integration

****Materials for Portfolio:****

- Executive summary (2-page visual overview of entire project)
- Selected visualizations (affinity map, comparison charts, key quotes)
- Methodology showcase (demonstrating research design skills)
- Before/after recommendations (Jamovi current state vs. proposed improvements)
- Reflection piece (what was learned about UX research and AI integration)

****Privacy-Preserving Presentation:****

- All participant quotes anonymized
- No identifiable information in visuals
- Aggregate data only; no individual profiles
- General demographics (e.g., "STEM professionals") rather than specifics

10. Quality Assurance

10.1 Validity Measures

****Internal Validity:****

- Standardized protocol for all participants (same tasks, same instructions)
- Pilot test with 1 participant to refine protocol before main study
- Consistent data collection instruments (SUS, NASA-TLX)
- Multiple data sources triangulated (performance metrics, questionnaires, interviews, observations)

****External Validity:****

- Diverse STEM disciplines increase generalizability within target population
- Remote testing mirrors real-world usage context (own computers, natural environments)
- Real statistical task (not artificial puzzle) enhances ecological validity

****Construct Validity:****

- Validated instruments (SUS, NASA-TLX) measure intended constructs
- Think-aloud protocol established method for capturing cognitive processes
- Heuristic evaluation grounded in Nielsen's established framework

10.2 Reliability Measures

****Inter-Rater Reliability:****
 - PURE study: 3 independent evaluators provide multiple perspectives
 - AI thematic analysis: 30 independent runs assess consistency of theme identification
 - Researcher's manual coding cross-validated with AI-generated themes

****Test-Retest Reliability:****
 - AI completes identical tasks 30 times; variability indicates task/interface ambiguity
 - Consistent findings across multiple participants suggest reliable problem discovery

****Instrument Reliability:****
 - SUS: Cronbach's alpha typically >0.90 (highly reliable)
 - NASA-TLX: Established psychometric properties

10.3 Audit Trail

****Documentation of Research Process:****
 - Version-controlled research plan (this document)
 - Detailed session logs with timestamps
 - Observer notes template filled for each session
 - Transcription verification notes
 - Analysis decision journal (coding choices, theme development rationale)
 - Prompt iteration log for AI studies

****Reproducibility:****
 - All prompts, code, and analysis scripts archived
 - Anonymized dataset available for verification (post-course, with participant consent)
 - Detailed methodology enables replication with different software or contexts

References

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24*(6), 574-594.

Bligård, L. O., & Osvalder, A. L. (2013). Predictive use error analysis—Development of AEA, SHERPA and PHEA to better predict, identify and present use errors. *International Journal of Industrial Ergonomics*, 43*(1), 43-57.

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. *The Cambridge Handbook of Expertise and Expert Performance*, 223-241.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data** (Rev. ed.). MIT Press.

Grier, R. A. (2015). How high is high? A meta-analysis of NASA-TLX global workload scores. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59*(1), 1727-1731.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. (1991). User interface evaluation in the real world: A comparison of four techniques. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 119-124.

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction** (2nd ed.). Morgan Kaufmann.

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods** (pp. 25-62). John Wiley & Sons.

Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, 206-213.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 249-256.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34*(4), 457-468.

Appendices

Appendix A: Test Materials Checklist

- [] cheese_nightmare_data.csv (100 rows, variables: cheese_type, nightmare_frequency)
- [] Task instruction sheet (PDF)
- [] Informed consent form
- [] SUS questionnaire (Microsoft Form)
- [] NASA-TLX questionnaire (Microsoft Form)
- [] Semi-structured interview guide
- [] Observer notes template
- [] Participant screening questionnaire
- [] PURE Standard Operating Procedure document (~30 steps)
- [] Heuristic evaluation reporting template
- [] AI prompts (task execution, heuristic evaluation, thematic analysis)

Appendix B: Participant Tracking Templates

****User Testing Tracker (Study 1):****
 [Spreadsheet with columns: ID, Recruitment Date, Screening Score, STEM Field, Session Date/Time,

Consent Signed, Recording Saved, Transcription Complete, SUS Score, NASA-TLX Score, Notes]

****PURE Evaluator Tracker (Study 2):****
 [Spreadsheet with columns: ID, UX Courses Completed, STEM Background, Session Date/Time, PURE Ratings Complete, Heuristic Eval Complete, Total Issues Identified, Severity Distribution]

****AI Trial Tracker (Study 3):****
 [Spreadsheet with columns: Trial ID, Timestamp, Task 1 Success, Task 2 Success, Task 3 Success, Task 4 Success, Total Time, Errors Logged, Notes]

Appendix C: Sample Consent Form

[Full text of adapted digital.gov consent template would be included here]

Appendix D: Data Management Folder Structure

```

\''\''
/UX604_Jamovi_Research/
  /00_Research_Plan/
    - Research_Plan_v1.0.pdf
    - Research_Plan_v1.1_Revised.pdf
  /01_Study_1_User_Testing/
    /Recordings/
      - P01_Session_Recording.mp4
    /Transcripts/
      - P01_Transcript.txt
    /Observer_Notes/
      - P01_Notes.pdf
    /Questionnaires/
      - SUS_Responses.csv
      - NASA-TLX_Responses.csv
      - Analysis_Study1.xlsx
      - Study1_Report.pdf
  /02_Study_2_PURE/
    /Recordings/
      - PURE_Ratings/
        - E01_Ratings.xlsx
    /Heuristic_Evaluations/
      - E01_Heuristic_Report.pdf
      - Analysis_Study2.xlsx
      - Study2_Report.pdf
  /03_Study_3_AI_Evaluation/
    /Trial_Logs/
      /Heuristic_Evaluations_AI/
      /Thematic_Analyses_AI/
      - Analysis_Study3.xlsx
      - Study3_Report.pdf
  /04_Final_Synthesis/
    - Cross_Method_Comparison.xlsx
    - Final_Recommendations.pdf
  /Appendices/
    - cheese_nightmare_data.csv
    - SOP_Document.pdf
    - All_Prom
  
```