A Usability Study of Jamovi for One-Way ANOVA

The Pufferfish of Statistics

LI Kanyu

04.03.2026

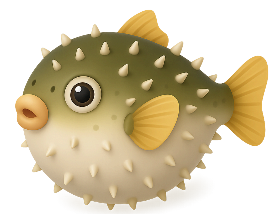
A Usability Study of Jamovi for One-Way ANOVA

The Pufferfish of Statistics

LI Kanyu

TL; DR

The pufferfish, or fugu, is the ultimate delicacy in Japanese cuisine, yet it is lethal if mishandled. The toxin isn't in the flesh itself, but in whether the chef knows where **not** to cut. Jamovi is to statistical analysis what the fugu is to Japanese cuisine: its sleek and elegant interface makes statistics unprecedentedly "palatable." However, precisely because it lowers the operational barrier, when a user lacks sufficient statistical knowledge, the system's silence becomes an invisible toxin. Users effortlessly proceed toward erroneous conclusions, completely unaware of the danger.



This study employed a mixed-methods design to evaluate Jamovi's usability in a one-way analysis of variance (ANOVA) workflow. Six participants with STEM backgrounds completed a full analysis process remotely, from data import to results interpretation. Data collected included standardized scale scores, think-aloud protocols, and behavioral observations.

The research revealed a core contradiction: participants reported extremely low subjective task load (NASA-TLX mean of 19.83/100), yet the System Usability Score was only 62.92 (below the industry benchmark of 70), and the six participants generated a total of 100 codable error events. Further analysis showed that over 70% of errors occurred during the data analysis phase (T3), with 83% of these being cognitive "Mistakes" rather than operational "Slips."

The most critical design flaw was this: when participants, due to insufficient statistical knowledge, selected an incorrect analysis variable (e.g., using an ID column as the dependent variable), Jamovi executed the analysis without any warning. This led participants to draw seemingly plausible research conclusions from invalid results. By simplifying its interface, Jamovi lowered the operational barrier but, due to a lack of feedback mechanisms and intelligent guidance, shifted the risk from "operational load" to "diagnostic load" and "decision-making risk." Within this framework, participants did not fail through struggle; they confidently and effortlessly arrived at the wrong destination.

The pufferfish is still delicious, but this kitchen needs better safety mechanisms. We recommend top priorities: (1) replacing the silent icon flicker with explicit error messages when variable types mismatch, and (3) Provide clear and comprehensive guidance for data QA, processing, and analysis to prevent mistakes as much as possible. These changes would allow less experienced chefs to cook with confidence.

Drafted Jan 18, 2026

Contents

Correspondence to
LI Kanyu
im@not.ci

Competing Interests
The authors declare no competing interests.

1. Objectives

1.a. Primary Research Question

Does Jamovi's interface design facilitate or hinder the accuracy of analysis when users perform a one-way ANOVA workflow?

1.b. Secondary Research Questions

- In the different stages of the statistical analysis workflow (data import, cleaning, analysis, visualization, interpretation), what specific usability issues arise?
- Is there a discrepancy between participants' subjective evaluations (perceived load, usability scores) and their objective performance (error rates, error types)?
- Is AI-assisted coding reliability validation a feasible and reliable method for qualitative analysis?

1.c. Research Background

The interface of statistical software is the primary point of interaction between a researcher and their data, and its design directly impacts the quality of the analysis. As an open-source alternative to SPSS, Jamovi aims to lower the barrier to statistical analysis with a more user-friendly interface. However, the tension between ease of use and analytical accuracy warrants scrutiny: does simplifying operations come at the cost of weakening necessary constraints on user decisions? This study aims to answer this question empirically and provide actionable design recommendations for interface improvement.

2. Methods

2.a. Participants

Six participants (P01–P06) were recruited. Inclusion criteria were: a STEM background, completion of a university-level statistics course, no professional UX experience, and less than 5 hours of experience with Jamovi. Participants' fields of study included medicine, education, economics, sociology, and computer science.

2.b. Experimental Task

Participants played the role of sleep researchers, using Jamovi to answer the question, "Does consuming different types of cheese before bed affect nightmare frequency?" The dataset was a machine-generated set of 100 random entries with defined distribution parameters (different means and variances), but the difference was not large enough to be detected by a conventional ANOVA, leading to a false negative result. The dataset contained three variables: `cheese_type` (4 levels: Cheddar, Brie, Swiss, Blue), `nightmare_frequency` (continuous), and a `ID` (an auto-incrementing integer sequence) for each row. Two data errors were pre-embedded for cleaning: one extreme outlier (200, should have been 2) and one formatting error ('3a', should have been 3). The latter would cause Jamovi to automatically classify the entire column as text, triggering a cascade of downstream issues.

The task sequence was as follows:

- **T1 Data Import & Verification** (approx. 5 minutes): Import the CSV file and confirm correct data loading.
- **T2 Data Cleaning** (approx. 5–8 minutes): Identify and handle the outlier and formatting error.

- **T3 Data Analysis** (approx. 8–12 minutes): Perform a one-way ANOVA to compare nightmare frequency across cheese types and interpret the statistical results.
- **T4 Data Visualization** (approx. 5–8 minutes): Create a chart to display the analysis results.

2.c. Data Collection

The test was conducted remotely via Google Meet, with participants operating Jamovi on their own computers while screen-sharing. The sessions were recorded automatically by Google Meet.

Quantitative Metrics:

System Usability Scale (SUS): A 10-item questionnaire with a score range of 0–100, where ≥70 is the industry-accepted benchmark (Bangor et al., 2008). Standard reverse scoring was used: for odd-numbered items, score = original score − 1; for even-numbered items, score = 5 − original score. The sum of these scores was multiplied by 2.5 to get the final score.

NASA Task Load Index (NASA-TLX): During scale creation, the researcher mistakenly omitted the "Performance" dimension, resulting in a 5-dimension scale (Mental Demand, Physical Demand, Temporal Demand, Effort, Frustration) instead of the standard 6. Upon realizing this error, we adopted a remedial analysis approach: we calculated the score using the Raw TLX method, taking the simple arithmetic mean of the 5 raw dimension ratings (on a 1–21 scale, lower is better), and then converted it to a 0–100 scale using the formula `(Raw - 1) / 20 × 100`. Due to the incomplete dimensions, the NASA-TLX scores from this study cannot be directly compared with standard norms; the reported scores serve only as a relative comparison among participants and a rough indicator of subjective load.

Qualitative Data:

- **Think-Aloud:** To capture participants' real-time cognitive processes.
- **Semi-structured interviews:** Post-test follow-up questions about the overall experience, difficult moments, and points of confusion.
- **Observational notes:** The researcher recorded key events and their timestamps.

2.d. Qualitative Data Analysis: AI-Assisted Coding Reliability Validation

This study employed an innovative method for coding reliability validation. The researcher first established a coding scheme(see Section 2.e). Subsequently, three large language models, DeepSeek v3 (DeepSeek-AI et al., 2024), Gemini 2.5 Pro (Comanici et al., 2025), and Grok 4.1 (xAI, 2025). Each independently coded the observational data for every participant 5 times, generating a total of 15 independent coding results.

The consistency metric used was the Simple Agreement Rate: for each coding dimension at each timestamp, we calculated the proportion of the most frequent code among all 15 coders. This was then averaged across all timestamps and dimensions. For any coding point with an agreement rate below 0.90, the researcher performed expert arbitration, and the final coding was based on this arbitrated version.

The overall coding agreement rates for the six participants were: P01 = 0.846, P02 = 0.888, P03 = 0.871, P04 = 0.856, P05 = 0.871, and P06 = 0.836. All exceeded the 0.80

threshold, indicating a very high level of consistency and reliability in the coding results.

2.e. *Coding Framework*

The following six-dimensional coding scheme was established. Each error event was coded across these dimensions (see full Codebook):

- **Error Type:** SLIP (correct intention, flawed execution), LAPSE (omission of a step), MISTAKE (wrong intention based on faulty understanding), NONE (successful operation, no error).
- **Error Subtype:** ACTION-SLIP, PERCEPTION-SLIP, OMISSION-LAPSE, MEMORY-LAPSE, RULE-MISTAKE, KNOWLEDGE-MISTAKE, MODEL-MISTAKE.
- **Usability Problem:** Multiple selections allowed. LACK-OF-FEEDBACK, POOR-AFFORDANCE, POOR-ERROR-MSG, INCONSISTENCY, BUG. Coded as NONE if the error was not caused by an interface design flaw.
- **Severity:** LOW (immediately recoverable, delay <30s), MEDIUM (required multiple attempts, delay 30s–2min), HIGH (task failure or required external help).
- **Task Phase:** T1 (Import), T2 (Cleaning), T3 (Analysis), T4 (Visualization).
- **Recovery Strategy:** IMMEDIATE, TRIAL-ERROR, STRATEGY-CHANGE, RESTART, HELP (required external assistance), CONTINUE (proceeded with the error).

2.f. *Quantitative Analysis Methods*

After coding, the error events were treated as a structured dataset for quantitative analysis.

Cluster Analysis: K-modes clustering (suitable for categorical variables) was used to group the error events, with the 6 coding dimensions (Error Type, Subtype, Usability Problem, Severity, Task, Recovery) as features. The optimal number of clusters, k was determined using the elbow method (testing k = 2–6). The seed was set to 42 for reproducibility.

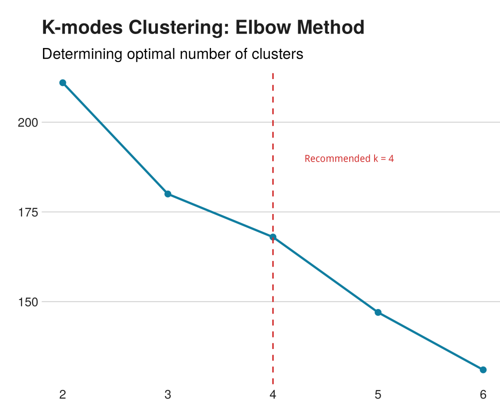


Figure 2: Elbow method plot showing the sum of within-cluster distances for different values of k.

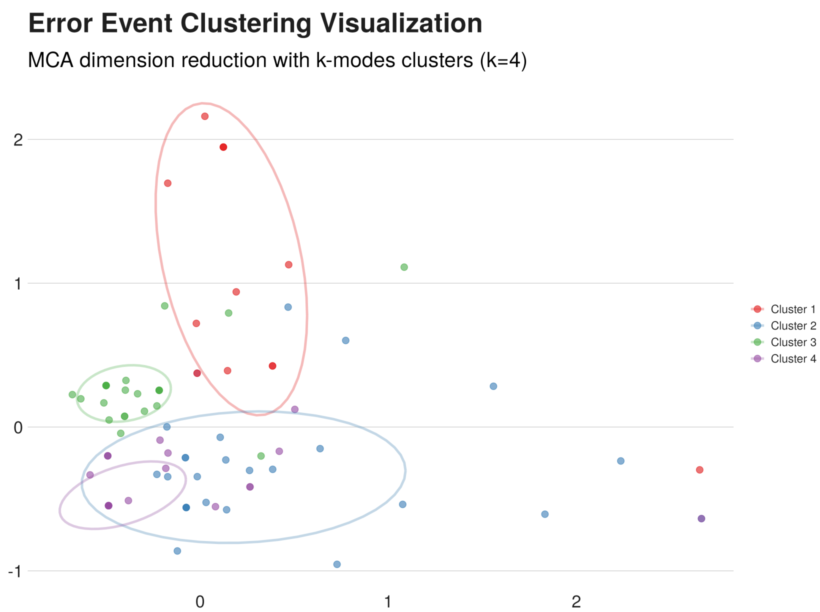


Figure 3: Two-dimensional MCA scatter plot showing the spatial distribution of the 4 clusters.

Statistical Testing: Permutation Tests were used to assess the statistical significance of key findings, with 10,000 iterations and a seed of 2026. Due to the limited sample size (n=100 error events) and low expected frequencies in some contingency tables, non-parametric methods were preferred over traditional chi-squared tests that rely on asymptotic distributions. Specifically:

For testing distribution dominance (e.g., "Is MISTAKE the predominant error type?"), a Bootstrap test was used: we resampled 10,000 times under the null hypothesis (uniform distribution) and calculated the probability of observing a proportion as extreme as the one found.

For testing associations between variables (e.g., "Are Error Type and Usability Problem related?"), a permutation test was used to obtain an exact p-value: we held the marginal distribution of one variable fixed, randomly permuted the other variable 10,000 times, and calculated the proportion of times the permuted chi-squared statistic exceeded the observed value.

3. Key Insights

3.a. Errors Were Concentrated in the Core Analysis Phase

Of the 100 error events, over 70% occurred during T3 (Data Analysis). A Bootstrap test indicated that the observed proportion of errors in T3 deviated from a uniform distribution ($p < 0.001$). A chi-square test yielded a consistent result ($\chi^2 = 47.79$, $p < 0.001$).

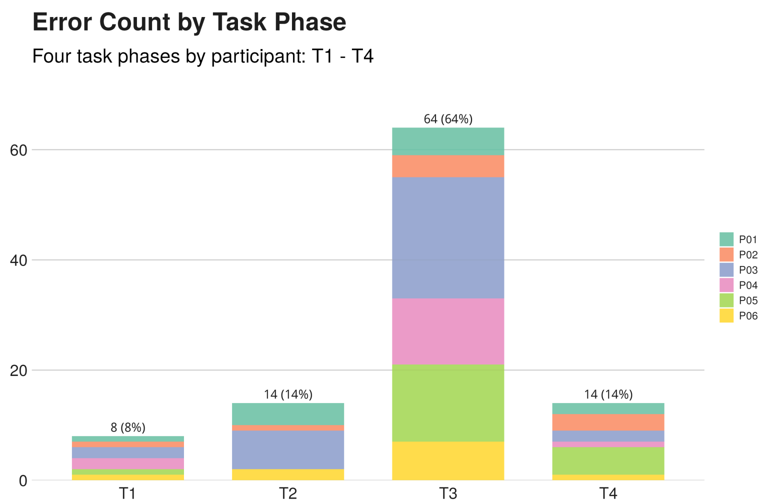


Figure 4: Distribution of error counts by task phase (stacked bar chart colored by participant).

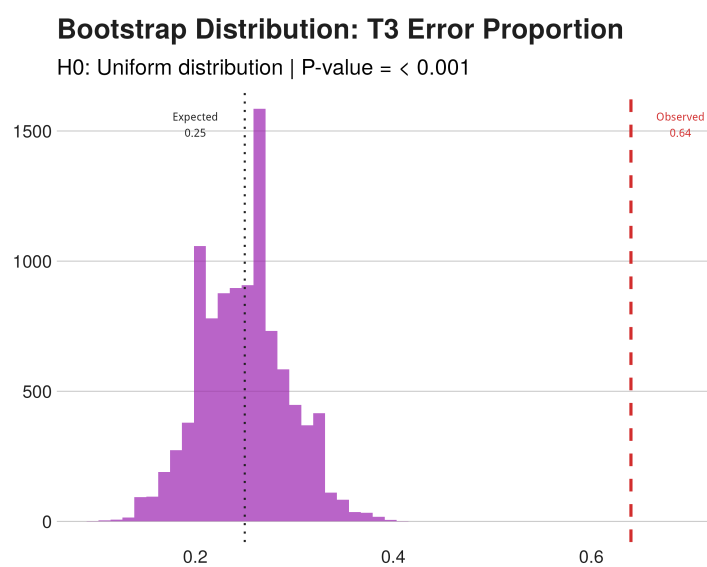


Figure 5: Comparison of the Bootstrap null distribution with the observed proportion of T3 errors.

Participants were able to complete data import (T1) and cleaning (T2) with relative ease, but encountered severe difficulties in T3, which required configuring analysis parameters, selecting the correct variables, and interpreting the output. Notably, several participants found data import to be the easiest part. P01 noted, “I could find it in three clicks, which matched my historical usage patterns,” and P03 agreed, “Having ANOVA and t-test right here as buttons… is much better than traditional software.” This suggests that while Jamovi successfully lowered the barrier to *finding* functions, it failed at the critical step of *using them correctly*.

Priority: Highest.

3.b. Cognitive Mistakes Dominated the Error Profile

Among all errors, 83% were classified as MISTAKEs, a proportion observed to deviate from a uniform null hypothesis by a Bootstrap test ($p < 0.001$).

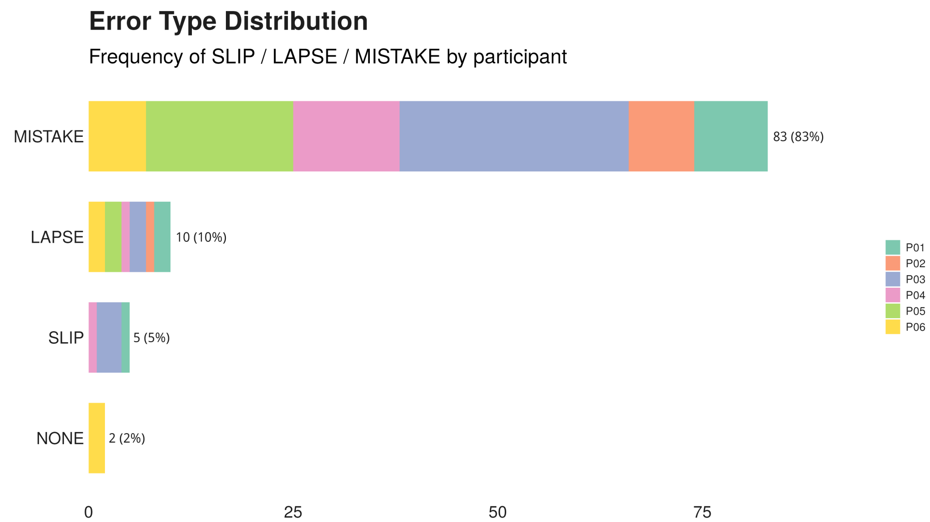


Figure 6: Distribution of error types (stacked bar chart colored by participant).

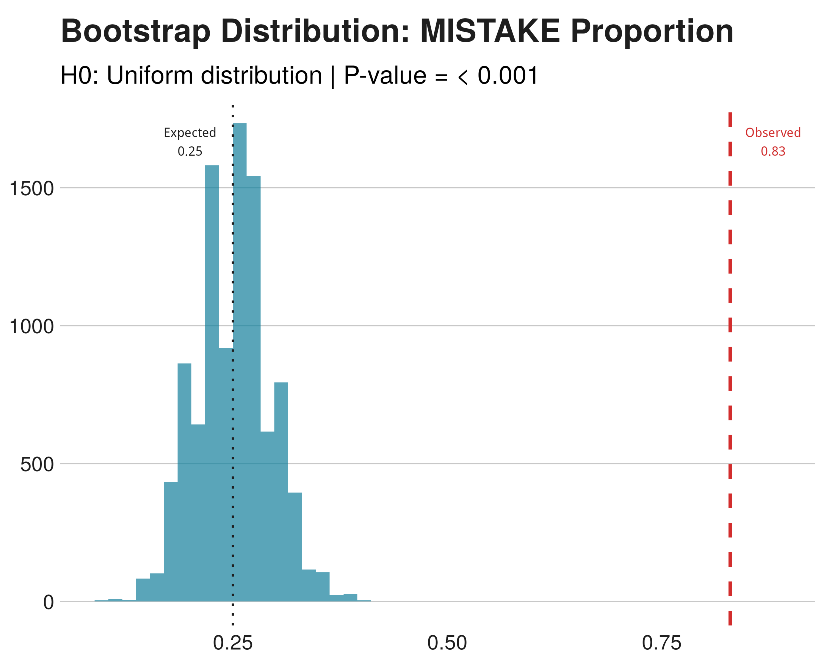


Figure 7: Comparison of the Bootstrap null distribution with the observed proportion of MISTAKEs.

This finding has profound design implications: the root cause of failure was not the size, position, or legibility of UI elements (which typically lead to SLIPs), but fundamental misunderstandings of statistical logic and system behavior. P02's experience is a prime example: his extensive experience with GraphPad became a liability. He admitted, "For someone used to GraphPad, suddenly switching to something like SPSS… is quite difficult." However, he also pointed out that Jamovi's automatic data type detection could be improved: "It could do simple keyword matching. Something like 'frequency' should be directly assigned as a continuous variable, instead of giving me… I didn't realize the default variable type was wrong until I couldn't plot the graph." Traditional interface tweaks (bigger buttons, clearer icons) cannot address this level of problem.

Priority: Highest.

3.c. Lack of Feedback Led to Inefficient Trial-and-Error Cycles (The Most Common Problem)

The largest error cluster identified by the cluster analysis was Cluster 4 (n = 33%), characterized by LACK-OF-FEEDBACK as the main usability problem and TRIAL-ERROR as the primary recovery strategy.

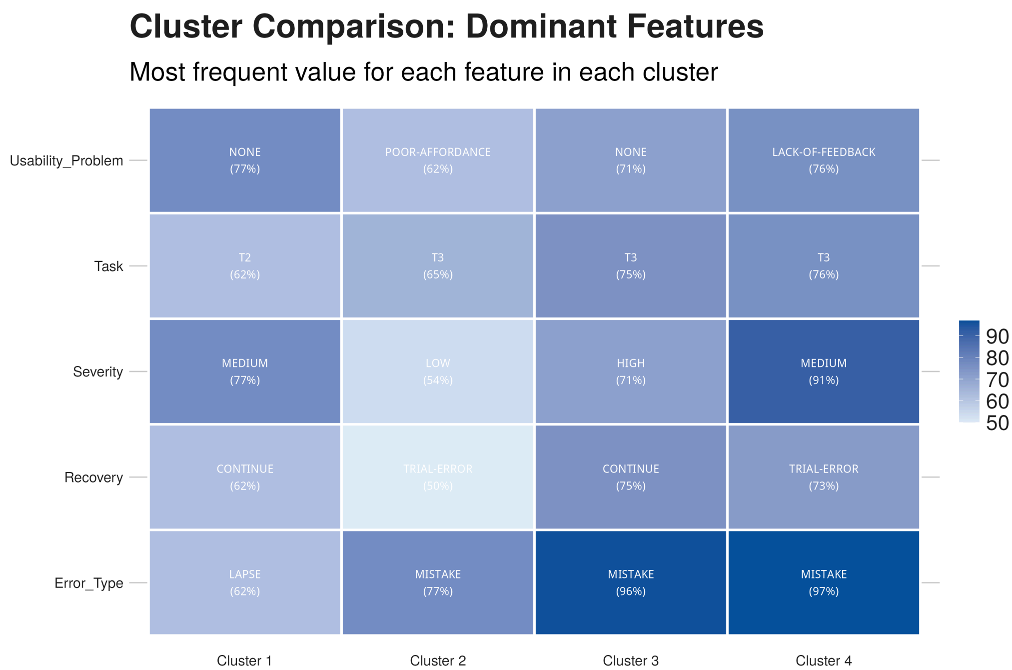


Figure 8: Heatmap comparing the dominant features of the four clusters.

A typical scenario was as follows: when a participant tried to drag a variable with a mismatched data type into an analysis box, the only feedback was a brief flicker of the icon in the target area. All six participants encountered this, but none understood its meaning. P06's description was the most direct: "When I couldn't add it, I was very confused. What does the flickering even mean? I knew it was supposed to go there, and it took me a minute or two to realize the variable setting might be wrong." P05 similarly reported that the "ruler icon was flickering" but had no idea what was happening. Observational records show participants repeatedly attempting the same action 3–5 times without being able to diagnose the failure.

Furthermore, this cluster included two related issues. First, the system silently handles non-compliant data during type conversion. After noticing the data type was text, P01 attempted a direct conversion, and the system automatically changed '3a'. The researcher later explained, "You forced a format conversion, and it turned '3a' into something else, so that data point wasn't cleaned." P02 responded, "I didn't notice that at the time," and suggested, "There should be a more obvious warning here." Second, when the analysis configuration was incomplete, the results panel simply showed an empty table frame with no explanation. P06 commented, “I feel that’s a bit of a problem.”

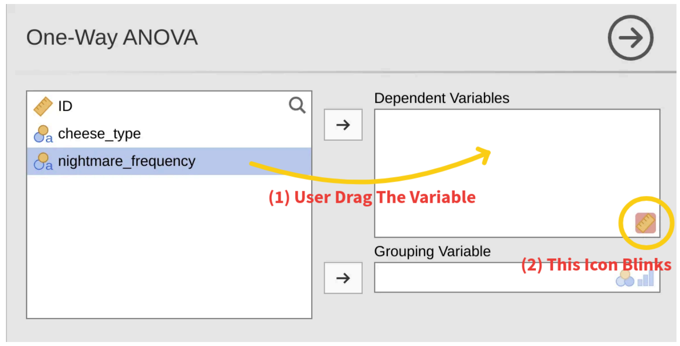


Figure 9: The screenshot that explains the issue.

Priority: High.

3.d. Knowledge-Based Errors Led to High-Stakes Misjudgments (The Most Dangerous Problem)

Cluster 3 (n = 28%) was the most severe cluster, characterized by KNOWLEDGE-MISTAKE as the main error subtype, HIGH severity, and CONTINUE (proceeding with the error) as the recovery strategy. The primary usability problem for this cluster was coded as NONE. This does not mean there was no design issue; rather, it indicates that the root cause was not a specific flaw in the interface itself, but the system’s failure to intervene when the user made a statistical error. It was an error of *omission*, not *commission*.

The core problem unfolded as follows: when participants were unable to drag the correct variable (`nightmare_frequency`) into the dependent variable box due to a data type mismatch, they resorted to selecting the only variable that “looked like it could be dragged in”—the row ID number. P04 articulated his logic: “I chose the ID as the dependent variable because, based on the icon in the bottom right, this field should contain content with that ruler icon.” P03 made the same choice based on the same icon-matching logic. The researcher later explained to P04, “What you analyzed was actually the data’s serial number… which is why it appeared significant.”

This reveals a critical user behavior pattern: lacking sufficient guidance, they attempt to find the “shortest path” rather than understanding the underlying meaning. They successfully executed the analysis and received a set of seemingly complete and credible statistical results. Based on this, participants drew firm but entirely incorrect research conclusions. P04 concluded, “There is no significant relationship between cheese type and nightmare frequency.” It is noteworthy that P06 was the only participant who did not fall into this trap—his language remained appropriately cautious: “Based on the current data, you cannot see any statistically significant effect. You would need more data, a control group.”

Furthermore, five out of the six participants drew incorrect conclusions. The researcher noted in a post-test interview, “You kept saying it wasn’t statistically significant, but you didn’t directly say there was no difference.” The other participants generally equated “$p > 0.05$” with “no effect,” a common misinterpretation in frequentist statistics (no evidence is not the evidence of no).

The misinterpretation of p-values by 5 out of 6 participants in this study is not an isolated phenomenon. A commentary published in Nature (Amrhein et al., 2019), co-authored by over 800 researchers, pointed out that a systematic analysis of 791 papers from 5 journals found that approximately 51% of the articles incorrectly equated "statistically insignificant" with "no effect" or "no difference." This means that the proportion of p-value misinterpretations observed in our micro-usability tests (83%, 5/6) is not a special case. This alignment suggests that p-value misinterpretations are not due to individual knowledge deficiencies, but rather a systemic problem caused by both statistical education and statistical tools.

Continuous tracking data from 2005 to 2019 shows that this problem has not improved over time, statistics courses, journal peer review, and industry standards have all failed to effectively curb it. This provides a rationale for the involvement of statistical software in improvement: when education and peer review cannot solve the problem on their own, intervention at the tool level is no longer "icing on the cake," but becomes a part worth exploring seriously.

This was the most dangerous finding, as it directly led to erroneous research conclusions, and the participants were completely unaware.

Priority: Highest.

3.e. Conflicting Conceptual Models Caused Operational Confusion

Cluster 2 (n = 26%) was characterized by MODEL-MISTAKE and POOR-AFFORDANCE. Participants frequently projected their usage habits from other software (especially Excel and GraphPad) onto Jamovi, leading to various operational errors.

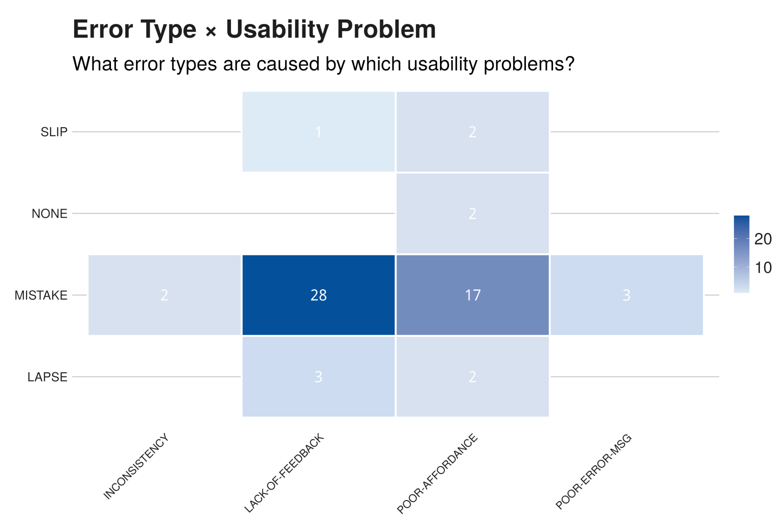


Figure 10: Heatmap of Error Type vs. Usability Problem.

P04 habitually copied and pasted analysis results into the data grid for side-by-side comparison, stating, "I want to paste this whole table next to the data to see it better." This is a standard operation in Excel but exposes a design issue in Jamovi where the boundary between the data area and the results area is unclear. P06 created a new ANOVA analysis item by clicking the menu, but he thought he was editing the existing one. He said, "I thought that when I clicked it, it was related to the one above, not that I was creating a new one." The researcher explained, "Each time you click One-Way ANOVA from the menu, you're actually creating a new analysis."

P05 discovered that the File menu in the macOS top menu bar lacked an "Open" function, violating platform design conventions. He noted, "First I'd think to click 'File' in my status bar, because on macOS you can usually import files from there, but then I saw there was no 'Open File' or anything like it." P06 criticized the translation quality of the multilingual interface: "Sometimes I can clearly feel it's a clunky translation… the Chinese translations for 'ordinal' and 'nominal' are a bit confusing… translating 'levels' to '水平' (shuǐpíng) is technically correct, but it's not appropriate here. '种类' (zhǒnglèi, meaning 'types' or 'categories') would be much better." P01 was even more direct: "I kind of want to switch it to English to understand it."

Despite these issues, several participants still expressed positive feedback about Jamovi. P01 found it "a bit better to use than Minitab, which we learned in class"; P02 said, "This software is much faster to get started with than SPSS"; P03 praised it, saying, "You can have ANOVA and t-tests as direct buttons here… which is so much better than traditional software"; and P06 concluded, "It's better than the others, but there might be a bit of a learning curve at the beginning."

Priority: Medium.

3.f. Lapses in Data Preparation Planted Hidden Dangers

Cluster 1 (n = 13%) was dominated by LAPSE and OMISSION-LAPSE errors, concentrated in T2 (Data Cleaning). Participants tended to quickly scan the data and proceed directly to analysis, or they only addressed the most obvious outliers.

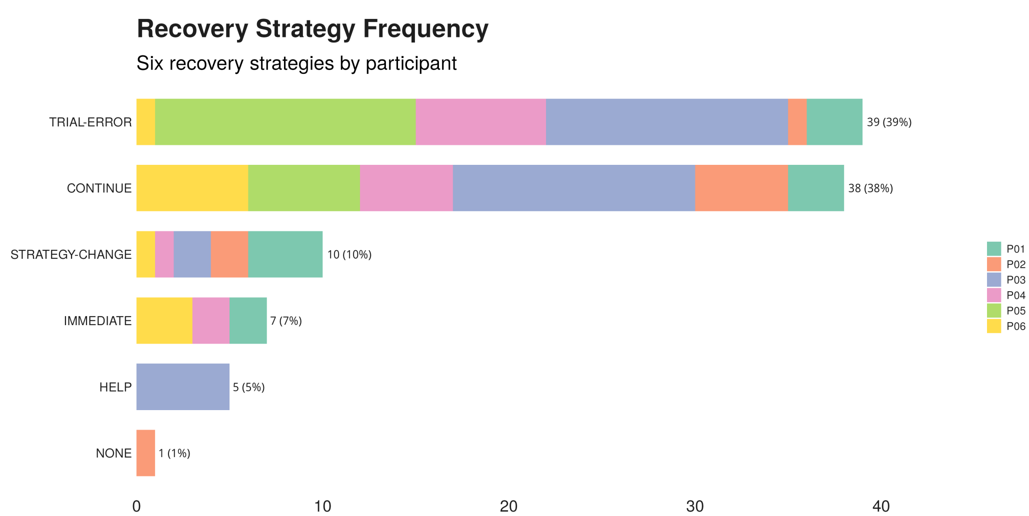


Figure 11: Frequency distribution of recovery strategies.

After deleting one of the two outliers (the value 200), P01 confidently stated, "Data cleaning, at a glance, I found one obvious problem and I've already deleted it… it's fine." P03 scanned the data table twice at the researcher's prompting but still failed to notice the '3a' formatting erro, even with guidance, P03 struggled to find it. P05 moved through the cleaning phase very quickly, skipping the check for '3a' altogether. In contrast, P04 was the only participant who identified and proactively deleted '3a', stating clearly, "This '3a' is obviously not data that should be here, so I'm deleting it."

The system's lack of a data quality check mechanism allowed flawed data to be silently carried into subsequent analyses. P02 suggested in a post-test review, "It could do simple keyword matching… maybe a more obvious warning would be good."

Priority: Medium.

3.g. Exploratory Patterns in Usability Problems and User Behavior

An exploratory question of this study was whether specific usability problems trigger specific user recovery strategies. For instance, does "lack of feedback" tend to result in "trial-and-error" behavior?

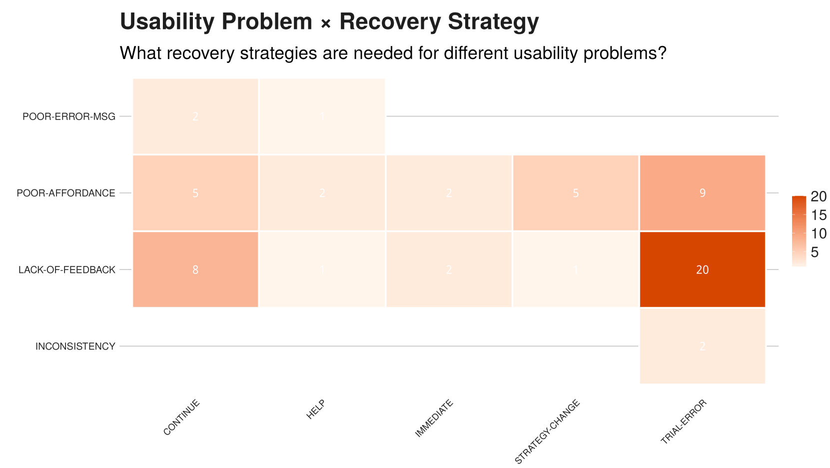


Figure 12: Heatmap of Usability Problem vs. Recovery Strategy.

Cross-tabulation and visual analysis of the data reveal observable concentrations in user behavior. The heatmap indicates specific co-occurrences, such as users frequently resorting to trial-and-error strategies when encountering a lack of system feedback.

To test whether these observed patterns represent systematic associations, we conducted permutation tests on two pairs of variables: (1) Error Type × Usability Problem, and (2) Usability Problem × Recovery Strategy. The analysis excluded instances where no specific usability problem was identified.

The permutation tests yielded $p = 0.177$ for the first pair and $p = 0.697$ for the second pair. Neither result reached the threshold for statistical significance.

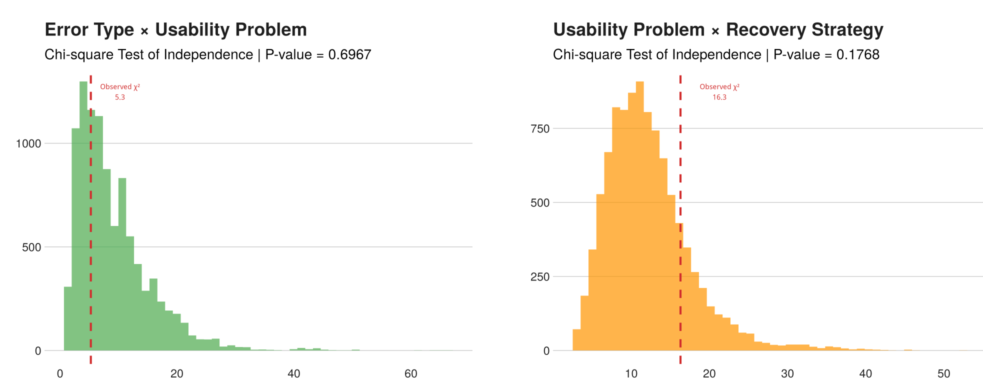


Figure 13: Permutation test distribution for variable associations.

However, the lack of statistical significance does not necessarily imply that the variables are entirely independent. Diagnostic checks of the contingency tables revealed numerous cells with low expected frequencies. Given the small sample size and the sparse distribution of events across multiple categories, the statistical tests lack the power required to distinguish potential underlying associations from random sample variation.

Consequently, the mappings between specific usability problems and user behaviors in this study must be interpreted as qualitative, exploratory patterns rather than statistically confirmed rules. For practical application, the failure to establish a statistically significant causal link does not invalidate the design insights. The observed patterns still provide actionable directions for improving the interface by directly addressing the identified heuristic violations.

3.h. *Results Summary*

3.h.i. *SUS and NASA-TLX:*

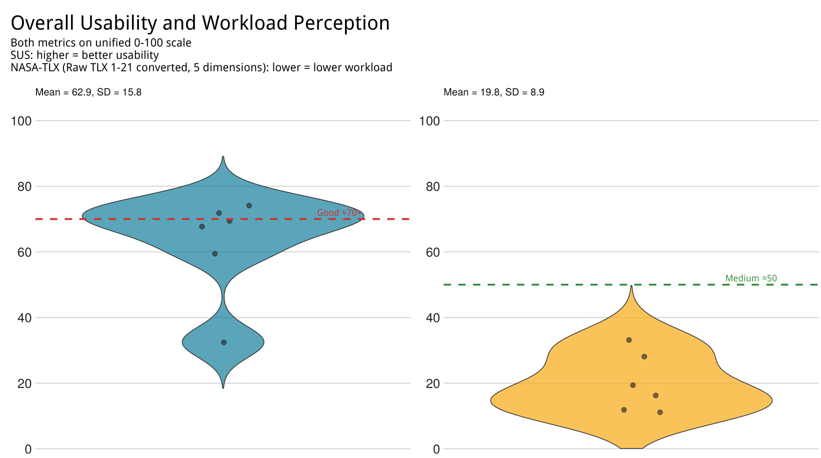


Figure 14: Violin plot comparing SUS and NASA-TLX scores (unified 0–100 scale).

Metric	Mean	SD	Range
SUS	62.92	15.77	32.5–75
Mod. NASA-TLX (0–100, lower is better)	19.83	8.89	12–33

The mean SUS score of 62.92 falls into the "marginal" usability range (Bangor et al., 2008), meaning about half the participants found the system difficult to use. In contrast, the mean NASA-TLX was only 19.83, indicating that all participants found the task to be very easy. Because the "Performance" dimension was missing, this TLX score cannot be directly compared to standard norms (e.g., the >54 high-load threshold proposed) (Grier, 2015), but its absolute level still suggests participants did not feel significant cognitive load.

This contradiction, "The system isn't great, but the task is effortless", points directly to the core finding of this study: Jamovi's problem is not that it makes users feel "tired" or "stuck," but that users make critical errors effortlessly, without any awareness. P06's comment perfectly captures this tension: "I feel like if I learned more, it would probably be much better to use." And P03, despite encountering significant statistical challenges, still commented, "Your usability is very good," confirming that low load and low accuracy can coexist.

3.h.ii. *Error Event Cluster Overview:*

#	Name	n	Primary Error Type	Primary Usability Issue	Severity	Primary Recovery
1	Lapse / Data Prep Trap	13	LAPSE	NONE	MEDIUM	CONTINUE
2	Model Mismatch / Poor Affordance	26	MISTAKE	POOR-AFFORDANCE	LOW	TRIAL-ERROR
3	Knowledge Error / High-Stakes Misjudgment	28	MISTAKE	NONE	HIGH	CONTINUE
4	Feedback Void / Trial-Error Loop	33	MISTAKE	LACK-OF-FEEDBACK	MEDIUM	TRIAL-ERROR

3.h.iii. *Statistical Test Results Summary:*

Finding	Method	Result
MISTAKE as dominant error type	Bootstrap	$p < 0.001$
LACK-OF-FEEDBACK + POOR-AFFORDANCE as dominant usability issue	Permutation	$p < 0.001$
T3 errors significantly higher than other phases	Bootstrap	$p < 0.001$
TRIAL-ERROR + CONTINUE as dominant recovery strategy	Permutation	$p < 0.001$
Association: Error Type × Usability Problem	Permutation	$p = 0.177$
Association: Usability Problem × Recovery Strategy	Permutation	$p = 0.697$

4. Actions

4.a. Balancing Simplicity and Guidance

Before discussing individual items, it is essential to highlight a structural tension that runs through all findings. Statistical software faces a unique challenge: statistical methodology itself is in constant evolution, and the combinatorial space of methods is vast. This means that even with significant UX resources dedicated to optimizing existing features, it is difficult to keep pace with the growth of domain knowledge. Jamovi is already one of the most user-friendly interfaces in its class, P01 called it "better than Minitab," P02 said it was "much faster to get started with than SPSS," and

P03 praised how “ANOVA is a direct button here… much better than traditional software.”

However, the core problem observed in this study was not that “the UI elements are bad,” but that “the system remains silent when the user needs help.” Five out of six participants effortlessly proceeded to wrong conclusions, and the system did nothing. This suggests that simple interface tweaks (bigger buttons, clearer icons) will not address the root cause. The system needs the ability to intervene at critical moments without compromising its operational simplicity.

In an informal post-session conversation, the researcher and a participant independently arrived at the same image: statistical software might need a kind of ever-present “intelligent assistant,” similar to Microsoft Office’s early Clippy, but tailored for statistical analysis. Clippy was removed from Office because modern office software operations have become standardized enough that an uninvited assistant is more of a nuisance than a help. But the complexity of statistical analysis is far beyond that of document editing. Tasks like method selection, assumption checking, and results interpretation will always involve judgments requiring domain knowledge. In such a context, a context-aware guidance mechanism may never be “superfluous.” This is, of course, a directional idea; its specific form should be determined by the design team based on technical feasibility and user research.

To extend the pufferfish analogy: fugu cuisine in Japan is safe not because the fish has become non-toxic, but because the industry has established a comprehensive safety system, from chef certification to standardized preparation procedures. The challenge for statistical software is similar: we cannot make statistics itself “non-toxic,” but we can build safety guardrails at the software level, allowing “chefs” of all skill levels to prepare data with greater confidence.

4.b. Operational Feedback and Safety Guardrails is the Highest Priority

How might we help users immediately understand the reason for an operational failure and the next steps to take? (Addresses Finding 3)

Current State: When a drag-and-drop action fails, the system’s only feedback is a brief icon flicker, which all six participants failed to understand. P06: “When I couldn’t add it, I was very confused. What does the flickering even mean?” P05 similarly reported the “ruler icon was flickering” but was unsure what happened. Participants repeated the same action 3–5 times without being able to diagnose the failure.

Participant Voice: P06 proactively suggested during the interview that the interface should provide “some kind of hint, maybe even a link to Wikipedia.” Users are already asking for richer, contextual guidance.

Design Precedent: Minitab’s Assistant feature provides a paradigm worth studying. The Minitab Assistant guides users through an interactive decision tree to select the correct statistical tool, provides on-demand definitions and visual examples for unfamiliar terms, and automatically generates a three-part report upon completion: a Summary Report (conclusions), a Diagnostic Report (outliers and test power), and a Report Card (warnings about assumptions and data quality). This “progressive disclosure” strategy embeds guidance at the very moment the user needs it, without cluttering the main interface.

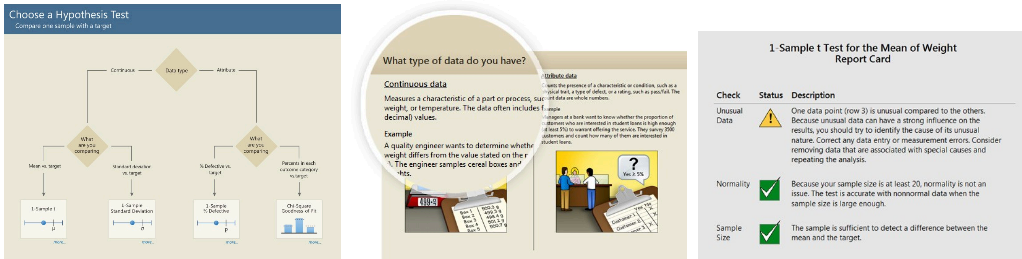


Figure 15: The guidance feature provided by Minitab

How might we make the consequences of data type conversion transparent to the user? (Addresses Findings 3 & 6)

Current State: During type conversion, '3a' was silently handled. The researcher later noted, "You forced a format conversion, and it changed '3a' to another value… this is clearly a bit of a system design problem." P02 later commented, "There should be a more obvious warning here."

One potential direction is to introduce a heuristic-based "data linter." Similar to a syntax checker in a code editor, but for statistical analysis configurations. For example, if over 95% of values in a column can be parsed as numbers, but a few contain non-numeric characters, the system can reasonably infer these are data entry errors rather than valid categorical labels. Even in 2026, manual data entry remains a reality in many research contexts (in fact, Jamovi itself supports it), so heuristic detection based on column content characteristics has broad practical value. Similarly, if a column's values range from 1–5 or 1–7, a wizard-style configuration screen could prompt, "This variable's characteristics are similar to a Likert scale," helping users make more appropriate variable type choices before analysis.

The "analysis wizard" concept proposed by P06 also points in this direction: "ask you questions step-by-step, and then produce the data." This type of guidance aligns with the classic Wizard design pattern from the Windows ecosystem: breaking down a complex task into a linear sequence of steps, presenting only relevant information at each stage to reduce cognitive load and configuration errors. Notably, Jamovi's interface already draws on many Windows and Office design elements, so introducing a Wizard pattern would be a natural extension rather than an awkward addition.

4.c. Data Preparation Flow and Discoverability

How might we make the structure of analysis results clear at a glance, making them easy to create, edit, retrieve, and manage? (Addresses Findings 3 & 5)

Current State: We observed several issues that appear different but share the same root cause. P06 accidentally created a new ANOVA instance after clicking the menu, saying, "I thought that after I clicked it once, it was related to the one above it, not that I was creating a new one." When an analysis configuration is incomplete, the results panel only shows an empty table frame with no explanation. P04 "accidentally deleted everything above" when right-clicking to delete a single analysis result; this was because his click landed in the outer area of the analysis item, which the system interpreted as selecting the entire results panel, causing the delete operation to affect all content. The crux of these issues is not a series of independent interaction flaws, but different manifestations of the same structural problem: the results panel lacks a clear organizational structure.

We understand Jamovi's design intent: to create a reporting area with a structure similar to a Jupyter Notebook, allowing users to freely insert notes and rich text between analysis results, completing the entire workflow from analysis to reporting in one place. This is an elegant design concept. However, cramming the creation, editing, management, and retrieval of results into a single, flat, scrolling view imposes a significant cognitive load on users as the number of analyses grows, as scrolling up and down is both tedious and prone to operational errors.

Design Precedent: SPSS's Output Viewer uses a dual-panel structure. On the left is a hierarchical tree-like outline (the Outline Pane), which organizes the output of each analysis into expandable/collapsible nodes. On the right is the actual content panel. Clicking any node in the left outline pane jumps to the corresponding result, and users can also reorder and delete results directly within the outline. This structure clearly separates "what I am looking at" from "how these things are organized," effectively solving the navigation problem as the number of results increases. Of course, Jamovi's Notebook-style design philosophy has fundamental differences from SPSS's traditional dual-panel approach, so a direct copy would not be appropriate.

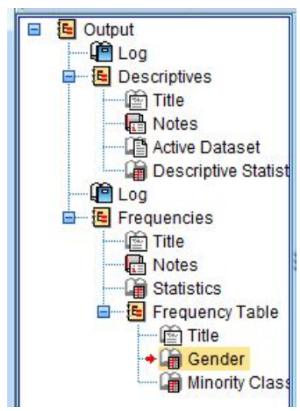


Figure 16: The result management of SPSS.

A potential direction that better aligns with Jamovi's existing design system is to overlay a lightweight navigation layer without disrupting the main Notebook-style view. The hideable table of contents design of Wikiwand (a third-party interface for Wikipedia) offers an interesting reference: when not interacting with the table of contents, it appears as a column of small dots floating on the right side of the page, with the dot corresponding to the current reading position highlighted. When the user hovers over or clicks this column of dots, the full table of contents panel expands, providing rich titles and navigation space. Similar floating table of contents (Floating TOC) patterns are quite mature in long-form content websites and document applications.

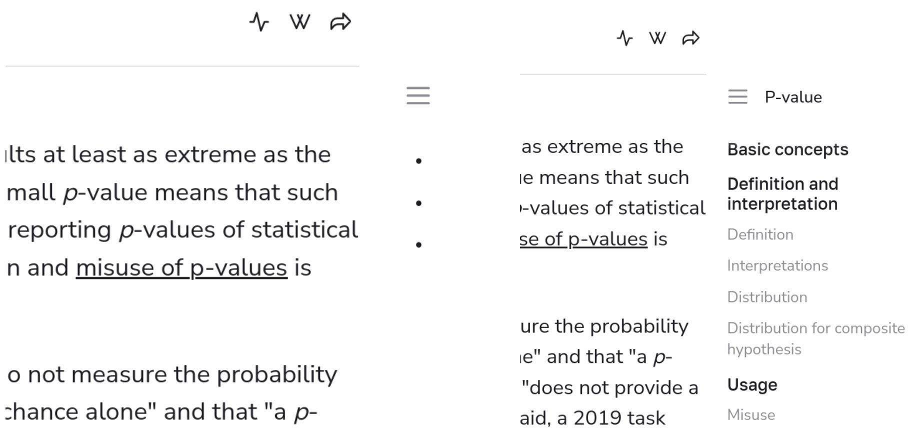


Figure 17: The ToC Design of WikiWand

For Jamovi, this pattern could solve multiple problems simultaneously. In its collapsed state, the series of dots itself serves as a lightweight index of results, allowing users to see at a glance "how many analyses I've done and which one I'm currently viewing." In its expanded state, the table of contents panel can house operations like renaming, deleting, and reordering results, and could even display warning messages. This would decouple the two needs, "freely writing in the Notebook" and "structurally managing analysis results", into different interaction layers.

How might we help non-English users feel confident, not confused, when using statistical terminology? (Addresses Finding 5)

Current State: P06 pointed out that "the Chinese translations for 'ordinal' and 'nominal' are a bit confusing" and suggested changing "水平" (levels) to the more intuitive "种类" (types). P01 said bluntly, "I kind of want to switch it to English to understand it." P06 also offered a concrete improvement: "even just a hyperlink below, that I can click to get a Wikipedia article," i.e., using progressive disclosure to provide context for terms without cluttering the interface.

5. Limitations

This study has several limitations. First, the sample size of six participants, while consistent with Nielsen's rule of thumb that 5±2 users can uncover 80% of usability problems (Nielsen, 2000), provides limited statistical power for quantitative inference, particularly in association tests. The sparse distribution of data points across categories restricts the capacity of statistical tests to confirm or reject systematic associations. Future research should use a larger sample to quantitatively evaluate these exploratory patterns.

Second, all participants were native Chinese speakers, and their perceptions of English statistical terms and Chinese translation quality may not apply to other language groups.

Third, the experimental task was a single one-way ANOVA workflow and cannot be generalized to Jamovi's other analysis modules.

Fourth, at the time of data collection, the Deepseek platform is conducting a gray-scale test for its new model version. This means that labeled data generated from the Deepseek website may originate from different model versions, representing a flaw in reproducibility. Subsequent research should note that when downloading original model weights, it is advisable to perform labeling tasks locally in a clean environment without vendor-customized system cards.

Fifth, the NASA-TLX scale omitted the "Performance" dimension due to an oversight, meaning its scores cannot be compared to standard norms.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

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. https://doi.org/10.1080/10447310802205776

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., & others. (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

DeepSeek-AI, Liu, A., Feng, B., Xue, B., & others. (2024). DeepSeek-V3 Technical Report. *Arxiv Preprint Arxiv:2412.19437*. https://arxiv.org/abs/2412.19437

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. https://doi.org/10.1177/1541931215591373

Nielsen, J. (2000). *Why You Only Need to Test with 5 Users* [Technical report]. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

xAI. (2025,). *Grok 4 Model Card*. https://data.x.ai/2025-08-20-grok-4-model-card.pdf

6. Supplemental Materials

6.a. AI Coding Consistency Analysis Detail

Consistency Analysis for P01

Timestamp	error_type	error_subtype	usability_problem	severity	task_phase	recovery_strategy	Average
01:48	0.71	0.57	0.50	0.50	1.00	0.43	0.62
02:42	1.00	1.00	1.00	0.70	1.00	1.00	0.95
02:59	0.50	0.50	0.83	1.00	0.67	0.67	0.69
03:14	1.00	1.00	0.40	0.73	1.00	0.93	0.84
03:25	1.00	1.00	0.80	1.00	1.00	0.60	0.90
04:01	1.00	1.00	0.85	0.85	1.00	0.85	0.92
04:26	1.00	1.00	0.93	0.87	0.87	0.73	0.90
05:00	0.86	0.57	1.00	1.00	1.00	0.86	0.88
05:10	1.00	1.00	0.86	0.57	1.00	0.43	0.81
06:24	1.00	0.89	0.56	1.00	1.00	0.78	0.87
06:33	1.00	0.75	1.00	0.75	1.00	1.00	0.92
07:50	1.00	0.50	1.00	1.00	1.00	0.50	0.83
10:01	1.00	1.00	0.53	0.60	1.00	0.67	0.80

10:53	1.00	1.00	0.86	1.00	1.00	0.57	0.90

Consistency Analysis for P02

Timestamp	error_type	error_subtype	usability_problem	severity	task_phase	recovery_strategy	Average
02:40	1.00	1.00	0.67	0.60	1.00	0.33	0.77
03:35	1.00	0.90	1.00	1.00	0.70	0.90	0.92
03:56	1.00	0.83	0.83	0.83	1.00	0.83	0.89
04:06	0.64	0.64	0.55	0.55	1.00	0.55	0.65
04:50	0.87	0.87	1.00	0.53	1.00	1.00	0.88
04:58	1.00	1.00	1.00	1.00	0.60	1.00	0.93
06:25	1.00	0.80	1.00	0.80	0.60	0.80	0.83
07:33	1.00	1.00	1.00	1.00	1.00	1.00	1.00
08:09	1.00	1.00	1.00	1.00	1.00	1.00	1.00
08:16	0.69	0.62	0.85	0.85	0.62	0.92	0.76
10:04	1.00	0.53	1.00	1.00	0.87	0.80	0.87
10:28	1.00	1.00	1.00	1.00	1.00	1.00	1.00
11:00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
12:00	1.00	1.00	1.00	1.00	1.00	0.50	0.92
Conclusion	1.00	1.00	1.00	0.50	1.00	1.00	0.92

Consistency Analysis for P03

Timestamp	error_type	error_subtype	usability_problem	severity	task_phase	recovery_strategy	Average
02:16	1.00	1.00	1.00	0.90	1.00	0.50	0.90
03:38	1.00	1.00	1.00	0.60	1.00	0.90	0.92
03:47	1.00	0.60	0.50	1.00	1.00	1.00	0.85
04:15	1.00	1.00	1.00	0.50	1.00	0.70	0.87
05:22	1.00	0.63	0.88	1.00	1.00	1.00	0.92
05:50	1.00	0.70	0.70	1.00	1.00	0.50	0.82
06:22	1.00	0.80	0.80	0.60	1.00	0.80	0.83
06:45	1.00	0.67	1.00	0.67	1.00	1.00	0.89
06:56	1.00	1.00	1.00	1.00	1.00	0.86	0.98
07:15	1.00	0.50	0.60	0.90	1.00	1.00	0.83
07:29	1.00	1.00	1.00	1.00	1.00	0.67	0.94
07:39	1.00	0.50	0.50	0.75	1.00	0.75	0.75
08:45	0.67	0.67	1.00	0.67	1.00	1.00	0.83
09:10	1.00	1.00	1.00	0.50	1.00	1.00	0.92
09:37	1.00	0.90	1.00	0.70	1.00	0.50	0.85
10:22	1.00	1.00	1.00	0.75	1.00	1.00	0.96
10:54	1.00	1.00	0.80	0.70	1.00	0.90	0.90
11:12	0.75	0.75	1.00	1.00	1.00	1.00	0.92
11:45	0.80	0.80	0.80	0.60	1.00	1.00	0.83
12:47	0.67	0.67	1.00	0.67	1.00	0.67	0.78
2:16	1.00	1.00	1.00	0.60	1.00	1.00	0.93
3:38	1.00	1.00	1.00	0.60	1.00	1.00	0.93

3:47	1.00	0.80	1.00	1.00	1.00	1.00	0.97
4:15	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5:22	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5:50	1.00	0.60	0.60	1.00	1.00	0.40	0.77
6:22	1.00	0.60	0.40	0.80	1.00	0.60	0.73
6:56	1.00	1.00	0.75	1.00	1.00	0.50	0.88
7:15	1.00	0.75	0.75	0.75	1.00	0.50	0.79
7:29	1.00	1.00	1.00	1.00	1.00	0.67	0.94
7:39	1.00	0.50	0.75	0.50	1.00	0.50	0.71
9:37	1.00	0.80	0.60	0.60	1.00	0.40	0.73

Consistency Analysis for P04

Timestamp	error_type	error_subtype	usability_problem	severity	task_phase	recovery_strategy	Average
04:38	0.69	0.46	0.62	0.92	1.00	0.38	0.68
08:17	1.00	1.00	1.00	1.00	1.00	1.00	1.00
08:45	1.00	1.00	0.92	0.77	1.00	0.92	0.94
09:10	1.00	0.58	1.00	0.75	1.00	0.92	0.88
09:13	0.92	0.38	0.62	0.85	1.00	0.77	0.76
09:27	0.82	0.36	1.00	0.64	1.00	0.36	0.70
10:28	1.00	0.60	0.47	1.00	1.00	0.87	0.82
12:04	1.00	0.73	0.33	0.87	0.93	0.60	0.74
13:22	1.00	0.88	0.50	0.75	0.75	0.63	0.75
14:08	1.00	0.90	0.70	1.00	0.90	0.40	0.82
15:10	1.00	1.00	0.57	1.00	1.00	1.00	0.93
4:38	1.00	1.00	1.00	1.00	1.00	0.50	0.92
8:45	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9:10	1.00	1.00	1.00	1.00	1.00	1.00	1.00
9:13	1.00	1.00	1.00	1.00	1.00	0.50	0.92

Consistency Analysis for P05

Timestamp	error_type	error_subtype	usability_problem	severity	task_phase	recovery_strategy	Average
02:42	1.00	1.00	1.00	0.63	1.00	0.63	0.88
03:04	1.00	1.00	1.00	1.00	1.00	1.00	1.00
03:27	1.00	1.00	1.00	1.00	1.00	1.00	1.00
04:16	1.00	1.00	1.00	1.00	1.00	1.00	1.00
05:27	0.67	0.67	1.00	0.67	1.00	0.67	0.78
05:41	1.00	1.00	1.00	0.71	1.00	0.64	0.89
06:00	1.00	0.67	1.00	0.83	1.00	0.75	0.88
06:30	1.00	0.83	1.00	0.75	1.00	0.67	0.88
06:34	1.00	1.00	0.67	0.83	1.00	0.67	0.86
07:01	1.00	0.50	0.42	0.75	1.00	0.67	0.72
08:15	1.00	1.00	1.00	0.75	1.00	0.50	0.88
11:12	1.00	1.00	1.00	0.80	1.00	0.60	0.90
13:45	0.56	0.56	1.00	0.78	1.00	0.67	0.76

14:10	0.93	0.87	0.67	0.93	1.00	0.73	0.86
14:44	1.00	0.40	0.60	0.60	1.00	0.80	0.73
16:31	1.00	1.00	1.00	0.75	1.00	0.75	0.92
17:30	0.75	0.75	1.00	0.75	1.00	0.75	0.83
18:07	0.58	0.25	0.83	0.92	0.92	1.00	0.75
18:08	1.00	1.00	1.00	0.50	1.00	1.00	0.92
18:30	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Consistency Analysis for P06

Timestamp	error_type	error_subtype	usability_problem	severity	task_phase	recovery_strategy	Average
03:12	0.57	0.43	1.00	0.64	1.00	0.93	0.76
05:06	0.83	0.67	0.50	0.50	1.00	0.50	0.67
05:36	1.00	1.00	1.00	1.00	1.00	1.00	1.00
07:03	0.73	0.40	0.87	0.73	1.00	0.67	0.73
08:05	0.43	0.43	0.71	0.57	0.57	0.57	0.55
08:21	1.00	1.00	0.50	0.83	0.83	0.83	0.83
12:24	0.91	0.73	1.00	0.55	1.00	0.55	0.79
12:38	1.00	1.00	1.00	1.00	1.00	0.67	0.94
12:52	0.50	0.50	1.00	1.00	1.00	1.00	0.83
13:56	0.67	0.67	0.67	0.67	1.00	0.67	0.72
14:07	0.75	0.75	0.75	1.00	1.00	1.00	0.88
14:30	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15:00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
CONCLUSION	1.00	1.00	1.00	1.00	1.00	1.00	1.00

6.b. *Manually Key Event Log*

6.b.i. *P01:*

01:31 The subject begins the task, attempting to find the file.
01:48 The user attempts to paste the file address into the search bar; the screen goes blank.
01:52 The user attempts to find the file using the browse button.
02:00 The user finds the file; Task 01 is complete.
02:42 After scanning the file, the user skips data cleaning and directly begins one-way ANOVA.
02:59 The native Chinese-speaking user is confused by the Chinese interface and requests an English interface.
03:14 The user attempts to drag the nightmare frequency data to the dependent variable, discovering a data type mismatch.
03:25 The user mistakenly drags data that was correctly dragged into the grouping variable back to the variable list.
03:38 The user begins enabling most statistical analysis functions in the ANOVA interface.
04:01 The user attempts to drag the variable into the dependent variable pane, failing due to data mismatch.
04:26 The user attempts to solve the problem through data transformation (incorrect

approach).
04:39 The user discovers that the data, which should have been continuous, is now in text format.
05:00 The user noticed the grid highlight box wasn't aligned with the grid data grid, but it was restored after dragging it twice.
05:10 The user tried directly changing the data to a numeric format (without filtering, incorrect approach).
05:30 The user found one of the two outliers and indicated they would temporarily set the issue aside.
05:51 The user correctly completed the variable settings.
06:00 The user began reviewing the report window.
06:13 The user discovered the significant impact of extreme values on the statistical results.
06:24 The user tried to find the automatic outlier removal function within the ANOVA function.
06:33 The user deleted one of the two outlier data points.
06:51 The user began reviewing the completion status of all tasks.
07:50 The user deleted the blank analysis results incorrectly created due to previous failed operations.
08:06 The user confirmed the output charts were correct.
08:30 The user began exploring other charting functions.
09:40 The user continued reviewing the analysis results.
09:56 The user attempted to create other charts.
10:01 The user attempted to create a scatter plot but incorrectly dragged the cheese type (name data) into the input. X-axis (continuous data)
10:27 User abandons scatter plot creation and deletes report
10:33 User opens "Plotting" and starts drawing box plot
10:48 User completes box plot
10:53 User attempts to draw a new plot (Pareto Plot) but indicates it doesn't make sense
Conclusion: The user indicates the results are not significant; the type of cheese consumed has no effect on whether or not one has nightmares (a classic statistical misinterpretation of p-values, but slightly corrected in subsequent discussion).

6.b.ii. *P02:*

02:10 The user drags the dataset into the window and begins data analysis.
02:30 The user opens the scatter plot function window and attempts to create a scatter plot.
02:40 The user tries to drag Cheese Type into the X-axis but fails due to a data type error.
02:51 The user attempts to analyze the results using descriptive statistics.
03:05 The user uses Nightmare Frequency as the analysis variable and splits it by Cheese Type.
03:35 The user notices that descriptive statistics cannot be performed due to a data type error.
03:56 The user finds the first outlier, 200, in the descriptive statistics table (which incorrectly classifies continuous values as categorical values).
04:06 The user is unsure if they have reversed "Split by" and Variables and tries to switch their positions.

04:50 The user attempts to draw a box plot, but due to an incorrect imported data type, no correct chart is output. The user is still unaware that the chart is incorrect.
04:58 The user begins to suspect a bug in the software.
05:56 The user starts searching for tools to automatically clean up outliers.
06:08 The user tries to identify outliers by drawing a scatterplot.
06:25 The user says they don't understand how to draw a scatterplot.
07:33 The user deletes the first outlier (an extremely large number).
07:48 The user successfully finds One Way ANOVA in the feature list and begins to determine whether to use the parametric or non-parametric version.
08:09 The user discovers a data type error.
08:16 The user correctly changes the data type, but still doesn't realize the outlier "3a" is formatted incorrectly.
09:35 The user opens the descriptive statistics tool, indicating they want to draw a box plot.
10:04 The user triggers a bug; they try to delete a report but mistakenly deletes all reports.
10:28 The user successfully draws the box plot and realizes that the previously incorrectly formatted data "3a" interfered with type resolution, causing the descriptive statistics and charting functions to malfunction.
11:00 The user restores the data using the correct procedure. ANOVA analysis results for bug removal
Conclusion: Users reported no significant difference; the type of cheese consumed had no effect on whether or not someone had nightmares (a classic statistical misinterpretation of p-values).

6.b.iii. *P03:*

2:16 The user begins the task and attempts to drag a file into the window. However, because the file is dragged into the "Statistical Results" pane instead of the "Data Table," data loading is not triggered. The user doesn't notice this and opens the File menu to attempt to import the data.
3:38 The user skipped data cleaning and went directly to the one-way ANOVA analysis interface.
3:47 The user attempted to drag the nightmare frequency data into the dependent variable several times, but failed due to data mismatch.
4:15 The user created an incorrect model and dragged the ID into the dependent variable according to the icon on the form.
5:02 The user created a new ANOVA analysis report.
5:22 The user tried again to drag the "nightmare frequency" data, which had a mismatched data type, into the dependent variable, but failed due to data type mismatch.
5:38 The user exited the one-way ANOVA interface; the results column currently shows that the data analysis is invalid (because the variables are not fully matched).
5:50 The user attempted to draw a scatter plot and put ID on the X-axis and nightmare frequency in the "Grouping" column (this was an incorrect operation because its semantics are continuous variables, not categorical variables).
6:10 The user reported "Unable to draw scatter plot".
6:22 The user tried to open the one-way ANOVA interface for the third time and attempted to drag ID... As the dependent variable, nightmare frequency was used as

the grouping variable (incorrect parameter configuration; only the type matched).
6:45 The user sighed, reiterating the error message "Insufficient observations" in the statistical report. The root cause was that too many numerical levels were incorrectly converted into nominal variables.
6:56 The user mistakenly opened the non-parametric version of one-way ANOVA.
7:15 The user mistakenly opened the independent samples t-test, using data ID as the dependent variable and nightmare frequency as the grouping variable, resulting in an incorrect analysis in the results column.
7:29 The user mistakenly opened the one-way ANOVA interface, dragging ID to the dependent variable (because they were the only visual cue for icon matching).
7:39 The user mistakenly opened the non-parametric version of one-way ANOVA again, using ID as the dependent variable and nightmare frequency as the grouping variable.
(The user abandons the task; the researcher provides clues, and the user continues trying)
8:45 The researcher prompts the user to perform data cleaning first.
9:10 The user closes the software filled with invalid reports and reopens a new software session.
9:23 The user successfully loads the data by dragging and dropping.
9:37 The user mistakenly attempts to use the "convert" function to convert the results of the nightmare data.
(9:56 The user abandons the task again)
10:11 The researcher reminds the user that data cleaning generally requires visual inspection.
10:22 The user reports seeing abnormal values; the researcher reminds the user to delete these abnormal values. The user completes the abnormal data deletion operation.
10:54 The user scans the entire data table twice and does not see any incorrectly entered data (the incorrectly entered data, the nightmare frequency, was entered as "3a").
11:12 The researcher reminds the user that there is a data point "3a," which is an incorrect entry.
11:45 Researchers explained to the user that an input error caused the system to misjudge the data type. The user immediately found the variable type editing interface in the right-click menu and changed the variable type.
12:47 The user completed the correct analysis.
(Ultimately, the user gave an incorrect statistical conclusion: there was no positive correlation between the type of cheese consumed and the frequency of nightmares.)

6.b.iv. *P04:*

4:38, the user started the operation, using the "Special Import" function, but selected the wrong file and imported a piece of JSON.
6:14 The user reopened the data file and saw the correct dataset.
6:58 The user found an excessively large value (200) and successfully deleted it.

7:27 The user found the erroneous value 3a and successfully deleted it.
8:17 The user noticed a value of 0, considered it reasonable, and correctly inferred.
8:23 The user opened the one-way ANOVA interface.
8:45 The user tried to drag Nightmare Frequency and Cheese type into the dependent variable, but they were both named variables in the system, and the data could not be dragged in successfully.
9:10 The user continued to try dragging Nightmare Frequency into the dependent variable, but only saw the icon flashing and did not understand its meaning.
9:13 The user mistakenly dragged Nightmare Frequency into the grouping variable.
9:27 The user clicked the arrow button to complete the analysis and found that the results panel only had a blank analysis table.
9:38 The user tried to open the "ANOVA" interface and repeatedly compared the differences between the ANOVA interface and the one-way ANOVA interface.
10:28 The user returned to the one-way ANOVA interface and mistakenly entered the data ID... The user dragged in the dependent variable (because it was the only variable with a matching data format) and tried dragging the nightmare frequency into the grouping variable (another incorrect operation).
11:51 The user deleted all records left during the previous exploration process, keeping only this incorrect one-way ANOVA result.
12:04 The user followed Excel's usage habits and copied the result to the data grid. However, this is not the "legitimate" way to use Jamovi, because the data grid is actually the data source, similar to SPSS. The user tried several times and successfully pasted all the data into the data grid.
12:52 The user activated several visualization options in the one-way ANOVA interface and generated a chart.
13:22 The user tried to "export" the analysis results to the data grid.
14:08 The user tried to "paste" the result chart into the data grid but found that the data grid does not support rich text format (of course, it doesn't).
15:10 The user started reporting the results.
The user performed ANOVA based on the data entry ID and Cheese type and concluded that there is a significant correlation between the type of cheese eaten before bed and the frequency of nightmares.

6.b.v. *P05:*

2:20 The user opened the software.
2:42 The user noticed that the File menu in the macOS top status bar lacked the Open function.
3:04 The user selected Open and used the Browser button to open the data.
3:27 The user discovered the abnormal data "3a".
4:16 The user correctly deleted this row of data.
4:58 The user opened the ANOVA menu.
5:27 The user hesitated between ANOVA, one-way ANOVA, and the nonparametric version and ultimately chose ANOVA (it can't be wrong; ANOVA can also be used for one-way ANOVA).
5:41 The user dragged the cheese type to the independent variable, but found that the data type mismatch prevented the operation from succeeding.
6:00 The user tried dragging cheest_type to the Dependent Variable field again, but it still failed.

6:30 The user tried dragging nightmare frequency into the dependent variable field, but found that the data type conflict prevented the dragging.
6:34 The user reported that "the ruler icon is flashing" but didn't know what was happening.
7:01 The user tried dragging the data ID into DV. (Because it's the only dependent variable)
7:08 The user dragged "cheese type" into Fixed Factors.
8:15 The user tried dragging "nightmare frequency" into the dependent variable column again.
11:12 The user opened the variable editing panel and adjusted the data type.
11:37 The user returned to the one-way ANOVA interface and correctly placed the variable in the corresponding area.
12:21 The user reported that the F-value was very small and said the results were not significant.
(The user gave the conclusion: "No significant effect")
13:45 The user tried to create a visualization using descriptive analysis (it didn't find some visualization options in the ANOVA panel).
14:10 The user mistakenly opened Pareto Plot for data visualization.
14:39 The user used descriptive statistics tools for data visualization.
14:44 The user mistakenly dragged "cheese_type" into the variable column and then tried dragging "nightmare frequency" into the grouping variable column, but failed.
15:18 The user found the Plot Tab at the top.
15:25 The user found the box plot.
15:49 The user correctly dragged "cheese" into the grouping column. Drag the nightmare frequency into the variable.
16:31 The user found a very large outlier, 200.
16:58 The user successfully eliminated another outlier.
17:30 The user reported the p-value and stated, "This does not prove that different varieties of cheese have no effect on nightmares (correct)."
18:07 The user drew a completely new box plot (but did not notice that the box plot automatically refreshes according to data changes).
(The user concluded: Different varieties of cheese have no significant effect on nightmare frequency; this is an incorrect conclusion. No evidence is not evidence of no.)

6.b.vi. *P06:*

3:00 User opens Jamovi to start the task.
3:12 Frustrated with Jamovi's built-in file manager, the user uses the "Browse" button to open the system's file manager.
3:36 User correctly imports the file.
4:03 User finds 3a invalid data and begins cleaning.
4:32 User finds 200 invalid data and begins cleaning.
5:06 The user's operating system is Japanese; finding it inconvenient to perform analysis in Japanese, the user switches to Chinese.
5:36 User saves the analysis results and restarts the software.
6:04 User reopens the interface.
6:15 User starts the one-way ANOVA task.
6:30 User begins to verify all functions provided by the analysis interface.

7:00 User enables all analysis functions.
7:03 User finds that only IDs can be dragged into the dependent variable interface.
8:05 User attempts to "insert" a column into the data grid; the variable metadata editing interface pops up.
8:21 User finds a data type problem and fixes the data type error.
8:49 User opens the ANOVA interface.
8:53 The user has successfully created an ANOVA analysis.
9:13 The user temporarily leaves the experimental environment.
10:20 The user returns to the experimental environment.
11:04 The user returns to the variable type interface to check.
12:24 The user, distrustful of the Chinese translation, switches the language back to English.
12:38 The user reopens the software.
12:43 The user reopens the previous analysis project.
12:52 The user creates a new ANOVA analysis entry.
13:36 The user opens the descriptive statistics tool to start plotting.
13:56 The user finds Density as a visualization tool for data patterns in descriptive statistics.
14:07 The user creates another new one-way ANOVA analysis entry.
14:14 The user activates the additional analysis functions in the one-way ANOVA tool.
14:23 The user activates the visualization tools within the one-way ANOVA tool.
(The user's final report states that there is no statistical evidence to support the hypothesis, the conclusion is correct, but the inference is incorrect. The user believes this is due to insufficient data volume, but the ground truth is that the variance of the sample generated by the original sample generation code is inherently large.)

6.c. AI Coding Prompt

Please Encode this data:

[KEY EVENT LIST PASTED HERE]

With the following codebook:

```
# Jamovi Usability Testing Codebook
## Coding Fields
Each timestamped event is coded with the following 9 fields:
| Field | Description | Example Value |
|------|-------------|---------------|
| `Timestamp` | Time stamp | `03:14` |
| `Description` | Event description | User attempts to drag data into dependent
variable... |
| `Error_Type` | User error type | `SLIP`, `LAPSE`, `MISTAKE` |
| `Error_Subtype` | Error subtype | `ACTION-SLIP`, `MEMORY-LAPSE`, `MODEL-
MISTAKE` |
| `Usability_Problem` | System issue | `POOR-AFFORDANCE`, `LACK-OF-FEEDBACK`,
`BUG` |
| `Severity` | Severity level | `LOW`, `MEDIUM`, `HIGH` |
| `Task` | Task phase | `T1`, `T2`, `T3`, `T4` |
| `Recovery` | Recovery strategy | `IMMEDIATE`, `TRIAL-ERROR`, `STRATEGY-
CHANGE`, `RESTART`, `HELP`, `CONTINUE` |
| `Notes` | Additional notes | Free text |
---
## 1. User Error Type (Error_Type)
```

```markdown
Based on Norman's (1988) classic classification.
### SLIP
- **Definition**: Correct goal, but execution error
- **Subtypes**:
  - `ACTION-SLIP`: Clicking error, dragging to wrong position
  - `PERCEPTION-SLIP`: Misreading, confusing similar elements
### LAPSE
- **Definition**: Forgetting to execute or omitting steps
- **Subtypes**:
  - `MEMORY-LAPSE`: Forgetting what has already been done
  - `OMISSION-LAPSE`: Completely skipping necessary steps
### MISTAKE
- **Definition**: Error in goal or understanding itself
- **Subtypes**:
  - `KNOWLEDGE-MISTAKE`: Lacking necessary knowledge (e.g., not understanding
data types)
  - `RULE-MISTAKE`: Using wrong method/strategy
  - `MODEL-MISTAKE`: Mental model inconsistent with the system
### NONE
- **Definition**: Successful operation, no error

## 2. Usability Problem (Usability_Problem)
Describes interface design defects (multiple selections allowed, separated by
commas).
| Code | Definition | Example |
|------|------------|---------|
| `POOR-AFFORDANCE` | Insufficient function visibility | Drag-and-drop target
area unclear |
| `LACK-OF-FEEDBACK` | Missing or untimely feedback | No prompt when operation
fails |
| `INCONSISTENCY` | Inconsistent interface | Different interaction methods
across modules |
| `POOR-ERROR-MSG` | Unclear error messages | Unhelpful error information |
| `BUG` | Software defect | Deleting one report deletes all |

## 3. Severity (Severity)
| Level | Definition | Judgment Criteria |
|-------|------------|-------------------|
| `LOW` | Minor issue | Immediately recoverable, delay < 30 seconds |
| `MEDIUM` | Moderate issue | Requires multiple attempts, delay 30 seconds-2
minutes |
| `HIGH` | Severe issue | Task failure or requires external assistance |

## 4. Task Phase (Task)
| Code | Task |
|------|------|
| `T1` | Data import and validation |
| `T2` | Data cleaning |
| `T3` | ANOVA analysis |
| `T4` | Data visualization |

## 5. Recovery Strategy (Recovery)
How users recover from errors (assessing interface fault tolerance).
| Code | Definition | Example |
|------|------------|---------|
| `IMMEDIATE` | Immediately recognize and correct | Immediately retry after
failed drag-and-drop |
```

`TRIAL-ERROR`	Find correct method through multiple attempts	Try multiple menus until function is found
`STRATEGY-CHANGE`	Change overall approach	Abandon conversion function, switch to manual modification
`RESTART`	Restart	Close and reopen software
`HELP`	Need external assistance	Researcher provides hints
`CONTINUE`	No recovery/Continue in error state	Continue subsequent tasks with error

Coding Process
1. **Read line by line** through original observation records
2. **Determine whether it is an error** (not all events are errors)
3. **Classify error type** (SLIP / LAPSE / MISTAKE)
4. **Identify system problems** (if any)
5. **Assess severity** (LOW / MEDIUM / HIGH)
6. **Mark task phase** (T1-T4)
7. **Record notes** (explain judgment basis or special circumstances)

Quality Check
After coding, verify:
- [ ] Each error has clear Error_Type and Error_Subtype
- [ ] Usability_Problem only filled when system defect exists
- [ ] Severity has judgment basis
- [ ] Statistical knowledge issues marked with `[Domain knowledge issue, not UX]`

Final coding result:

```
{
  "metadata": {
    "coder_name": "string",
    "participant_id": "string",
    "coding_date": "YYMMDD-HHMM",
    "software_target": "Jamovi",
    "version": "1.0"
  },
  "coding_records": [
    {
      "timestamp": "MM:SS",
      "original_record": "string",
      "coding_fields": {
        "error_type": "SLIP | LAPSE | MISTAKE",
        "error_subtype": "ACTION-SLIP | PERCEPTION-SLIP | MEMORY-LAPSE | OMISSION-LAPSE | KNOWLEDGE-MISTAKE | RULE-MISTAKE | MODEL-MISTAKE",
        "usability_problem": [
          "POOR-AFFORDANCE",
          "LACK-OF-FEEDBACK",
          "INCONSISTENCY",
          "POOR-ERROR-MSG",
          "BUG"
        ],
        "severity": "LOW | MEDIUM | HIGH",
        "task_phase": "T1 | T2 | T3 | T4",
        "recovery_strategy": "IMMEDIATE | TRIAL-ERROR | STRATEGY-CHANGE | RESTART | HELP"
      },
      "rationale": {
        "decision_basis": "string",
```

```
        "notes": "string"
      }
    }
  ]
}
NOTICE THAT IF A LINE IS NOT RECORDING AN FAULT OPERATION, YOU CAN SKIP IT,
DON'T ENCODE ANYTHING.
```

Give JSON directly, don't say anything else.

7. Procedural materials

7.a. Consent Form

Purpose of the Study

This study aims to investigate the user experience (UX) of the open-source statistical software, Jamovi. By collecting feedback on its user interface, functional logic, and visual feedback, we hope to identify the software's strengths and weaknesses. The findings will provide insights for the future optimization of open-source statistical tools.

Procedures

If you decide to participate, you will be asked to:

- Complete an online questionnaire about your habits using Jamovi (approximately 10-15 minutes).
- Alternatively (for usability testing): Perform several specific statistical tasks under observation while sharing your thoughts aloud (approximately 30 minutes).

This study does not involve any tasks that are physically or psychologically stressful.

Risks and Discomforts

The risks to participants in this study are **Minimal Risk**. You may experience mild visual fatigue from using the software or brief anxiety due to unfamiliarity with certain tasks. You may request a break or stop your participation at any time.

Potential Benefits

You may not directly benefit from your participation in this study. However, your feedback will contribute to the improvement of open-source scientific tools, making statistics more accessible and user-friendly for students and researchers worldwide.

Conflict of Interest Statement

The researcher, Li Kanyu, and supervisor, Susie Simon, have no financial or vested interests in the official Jamovi team or its commercial competitors. This study is conducted for purely academic purposes.

Confidentiality and Data Management

- **Anonymity:** All collected data will be anonymized. Your name will not be linked to the research findings.

- **Data Storage:** Data will be stored in an encrypted format on a secure Wilfrid Laurier University (WLU) cloud server or a password-protected computer.
- **Retention Period:** Raw data will be retained for a period of [e.g., 5 years] after the study's completion or publication, after which it will be permanently deleted.

Voluntary Participation

Participation in this study is completely voluntary. You may withdraw from the study at any time, for any reason, without penalty. If you choose to withdraw, any data you have provided will be destroyed.

Contact Information

- **For questions about the study,** please contact Li Kanyu (lixx@mylaurier.ca) or Supervisor Susie Simon (ssimon@wlu.ca).
- **For questions about your rights as a research participant,** please contact:
    1. WLU Research Ethics Board
    2. Phone: 548.889.3518
    3. Email: REB@wlu.ca

Consent Statement

By clicking "Agree" or signing below, you confirm that you have read and understood the information above, are 18 years of age or older, and voluntarily consent to participate in this study.

7.b. Facilitator Guide

Material Verification:

- **Test Dataset:** `cheese_nightmare_data.csv`.
- **Hardware/Software:** Zoom (with Cloud Recording), participant's PC with Jamovi pre-installed.
- **Online Forms:** Links for SUS and NASA-TLX questionnaires.

Introduction and Ethics (5 Minutes)

- **Welcome:** Thank you for participating in this Jamovi statistical software usability study.
- **Informed Consent:** - Confirm electronic consent form is signed.
    - Inform that audio, video, and screen sharing will be recorded.
- **Rights Statement:** Participants may withdraw or skip tasks at any time without providing reasons.
- **Think-Aloud Training:**
    - **Instruction:** "Please speak your thoughts, feelings, or confusion aloud. Do not just tell us what you are doing; tell us why you are doing it."
    - **Demonstration:** Brief demo of searching for a computer folder while thinking aloud.

Task Execution Phase (20-30 Minutes)

Research Scenario: You are a sleep researcher investigating whether eating cheese before bed leads to nightmares.

Task 1: Data Import and Verification

- **Instruction:** "Import `cheese_nightmare_data.csv` into Jamovi and verify if the data loaded correctly."
- **Observation:** Can the participant locate the import function? Do they confirm the 100 rows and 2 columns structure?

Task 2: Data Cleaning

- **Instruction:** "Check the dataset for errors or outliers and clean them based on your judgment."
- **Hidden Challenges:** One extreme outlier (input error) and one formatting error (letters mixed in a numeric field).
- **Moderator Intervention:** Do not provide success criteria. If the participant is silent, prompt: "What are you thinking right now?"

Task 3: One-Way ANOVA Execution

- **Instruction:** "Perform a One-Way ANOVA to compare nightmare frequency across different cheese types and provide a statistical interpretation."
- **Observation:** Accuracy of variable selection (DV/IV) and understanding of the output tables.

Task 4: Data Visualization

- **Instruction:** "Create a suitable chart showing the relationship between cheese type and nightmare incidence."
- **Success Criteria:** Selection of appropriate chart type (e.g., bar chart or boxplot) with interpretable formatting.

Subjective Scale Assessment (5 Minutes)

Participants complete the following via provided links:

- **SUS (System Usability Scale):** 10 items assessing overall software usability.
- **NASA-TLX (Task Load Index):** 6 dimensions assessing cognitive workload during tasks.

Semi-Structured Interview (10 Minutes)

- **General Experience:** How would you describe your experience using Jamovi for analysis?
- **Ease of Use:** Which task was the easiest? Why?
- **Pain Points:** Which task was the most difficult or frustrating? What were the specific challenges?
- **Critical Incident Review:** "I noticed a long pause during [specific action]; what were you thinking then?"

Conclusion and Data Handling (2-3 Minutes)

- Stop recording and save files.
- **Notification:** All identity information will be anonymized (IDs P01-P06). Recorded data will be destroyed by April 15, 2026.

7.c. Participant Tracker

Start Date	Consent Agree	Participant ID	sus_01	sus_02	sus_03	sus_04	sus_05	sus_06	sus_07	sus_08	sus_09	sus_10
2/4/26 20:17	Agree	P01	Agree	Disagree	Agree	Disagree	Disagree	Disagree	Agree	Disagree	Neurtal	Disagree
2/4/26 21:15	Agree	P02	Neurtal	Neurtal	Disagree	Agree	Neurtal	Agree	Disagree	Neurtal	VeryDisagree	Agree
2/5/26 7:20	Agree	P03	Agree	Agree	VeryAgree	Neurtal	Agree	Disagree	VeryAgree	VeryDisagree	Agree	Disagree
2/5/26 20:21	Agree	P04	Agree	Disagree	Agree	Disagree	Agree	Agree	VeryAgree	VeryDisagree	Agree	Agree
2/5/26 21:06	Agree	P05	Neurtal	Disagree	Neurtal	Disagree	Neurtal	Disagree	Agree	Disagree	Agree	VeryAgree
2/5/26 10:32	Agree	P06	Agree	Disagree	Agree	Disagree	Agree	VeryDisagree	Neurtal	Agree	Agree	VeryDisagree

Participant ID	tlx_mental_demand	tlx_physical_demand	tlx_temporal_demand	tlx_effort	tlx_frustration
P01	10	2	2	7	3
P02	8	1	2	6	4
P03	10	15	3	4	1
P04	6	2	2	5	2
P05	8	2	7	10	11
P06	4	1	2	5	4

Participant	Link
P01	https://drive.google.com/file/d/1kXNGS4LGgJZkclhnzOhg1X_LiC-Qlf_J/view?usp=drive_link
P02	https://drive.google.com/file/d/1SHX_h5XpWqA4VMwABBPaW27yrUB0S2JV/view?usp=drive_link
P03	https://drive.google.com/file/d/1H6Ms0VIxMDCXhno2yyHW-e6yEqsVmD3k/view?usp=drive_link
P04	https://drive.google.com/file/d/1E7n9kmCEqy2HrodHwmUu8P6cb1xrCkIi/view?usp=drive_link
P05	https://drive.google.com/file/d/179M8gNOTG_gsJmLPayJeZ3UVGHIU6Ht0/view?usp=drive_link
P06	https://drive.google.com/file/d/1rDH56b3useckJOe25rdG_6LOB4Mf6M9O/view?usp=drive_link