# Wason Selection Task Experiment: Evaluation

Cognitive Systems Project

January 26, 2026

## 1 Statistical Evaluation Plan

### 1.1 Overview of Design

The experiment uses a between-subjects, single-factor design with one trial per participant. Participants are randomly assigned to one of three framing conditions:

1. **Abstract context (A)**: numbers and colors
2. **Familiar real-world context (B)**: age and alcohol
3. **Unfamiliar structured context (C)**: rule-based access/permission scenario

The logical structure of the Wason Selection Task (WST) is identical across conditions; only the semantic framing differs. Dependent variables are collected per participant, including correctness, response time, selection behavior, and confidence.

### 1.2 Primary Hypothesis

The primary hypothesis concerns performance differences across contextual framing conditions:

$$\text{Accuracy}_B > \text{Accuracy}_C > \text{Accuracy}_A \tag{1}$$

The most important planned comparison is expected to be the difference between the familiar and abstract conditions:

$$\text{Accuracy}_B > \text{Accuracy}_A \tag{2}$$

Differences between familiar and unfamiliar structured contexts are treated as exploratory, conditional on whether the unfamiliar structured scenario supports permission-based reasoning.

### 1.3 Ideal Inferential Tests (Final Analysis)

#### 1.3.1 Correctness (Primary DV)

Correctness is binary (correct vs. incorrect). The most appropriate final analysis is logistic regression:

$$\text{Correct} \sim \text{Condition} \tag{3}$$

This model yields an omnibus test of whether correctness differs across the three conditions. The primary effect of interest is tested using planned contrasts:

- **Primary planned contrast:** Familiar (B) vs. Abstract (A)
- **Secondary contrasts (exploratory):** Familiar (B) vs. Unfamiliar structured (C), and Unfamiliar structured (C) vs. Abstract (A)

Effect sizes will be reported as odds ratios (OR) with 95% confidence intervals. In the presence of sparse data or complete separation, penalized methods (e.g., Firth logistic regression) will be explored.

Emergency continuity plan: Fisher's exact tests (pairwise) when expected counts are small.

**Effect Size Reporting for Accuracy.** Multiple effect sizes are appropriate:

- **Difference in proportions:** $\Delta p = p_1 - p_2$
- **Odds ratio:** Odds ratios (OR) from logistic regression

### 1.3.2 Time to Submission

Completion time is continuous and typically right-skewed in browser experiments. The ideal analysis uses either:

- linear regression / ANOVA on log-transformed time,

$$\log(\text{Time}) \sim \text{Condition} \tag{4}$$

### 1.3.3 Selection Changes

Selection changes are non-negative integer counts. A generalized linear model should be suitable:

$$\text{Changes} \sim \text{Condition} \tag{5}$$

using a Poisson model.

### 1.3.4 First Card Selected

The first selected card is categorical. A contingency table analysis (Condition $\times$ FirstCard) is appropriate:

- chi-square test for independence

### 1.3.5 Confidence Ratings

Confidence is recorded on a 0–100 scale. The ideal analysis is linear regression / ANOVA:

$$\text{Confidence} \sim \text{Condition} \tag{6}$$

Because confidence may be strongly related to correctness, an additional model can be informative:

$$\text{Confidence} \sim \text{Condition} + \text{Correct} \tag{7}$$

# 2 Current Status: Preliminary Results (Pilot Data)

## 2.1 Dataset Summary

At the time of this interim analysis, the database contains 12 valid participant sessions distributed as follows:

- Abstract context (A): $n = 4$
- Familiar context (B): $n = 6$
- Unfamiliar structured context (C): $n = 2$

Group sizes are notably imbalanced and the unfamiliar structured condition has very low $n$, making inferential statistics unreliable at this stage. Therefore, current results are interpreted descriptively.

## 2.2 Accuracy by Condition

Observed correctness rates are:

- Abstract (A): $2/4 = 50\%$
- Familiar (B): $3/6 = 50\%$
- Unfamiliar structured (C): $0/2 = 0\%$

Zuerst Deskriptiv. e.g. Balken diagramm Deskriptiv verschaulichen wie viele sind auf Modus Tollens, pollens, yada yada drauf eingefallen