Power Analysis: AI Investment Decision Experiment

Within-subjects design with 6 conditions

AUTHORS
Jason Collins
Iñigo De Juan Razquin
Jianhua Li

PUBLISHED
June 7, 2025

1 Overview

This document presents a power analysis for a within-subjects experiment investigating how different methods of delivering AI explanations affect human-AI collaborative decision-making. The analysis addresses the question:

> How many participants do we need to reliably detect a 3 percentage point improvement in investment performance when using alternative AI explanation delivery methods?

1.1 Research Context

While AI can provide valuable decision support, users often fail to engage meaningfully with AI explanations, limiting the potential for complementary human-AI performance. This experiment tests whether alternative explanation delivery methods can promote deeper engagement and better decision-making.

Specifically, we test three interventions designed to encourage more deliberate engagement with AI explanations:

1. **Request**: Users must actively click to access the AI recommendation and explanation
2. **Update**: Users make an initial decision, then can revise it after seeing the AI's input
3. **Wait**: The explanation is provided after a deliberate delay

The "Explanation" condition (immediate explanation) serves as the control against which we compare the three intervention strategies. The "No AI" and "No Explanation" conditions provide performance benchmarks.

1.2 Design

- **Within-subjects design**: Each participant experiences all 6 conditions in random order, serving as their own control
- **6 experimental conditions**:
  - No AI (benchmark)
  - AI without explanation (benchmark)
  - AI with immediate explanation (control)
  - Three explanation delivery interventions (request, update, wait)
- **36 decisions per participant**: 6 investment decisions × 6 conditions
- **Realistic effect size**: 3 percentage point improvement (75% → 78% success rate)
- **Multiple testing correction**: Bonferroni adjustment for 5 comparisons vs control

1.3 Analysis Approach

We employ three statistical methods to obtain power estimates:

1. **Paired t-tests**: Simple comparison of participant scores between conditions
2. **Generalised Linear Mixed Models (GLMM)**: Model individual decisions with random effects
3. **Generalised Estimating Equations (GEE)**: Population-level marginal effects

Each method handles the repeated-measures structure differently, providing converging evidence for sample size requirements.

1.4 Assumptions

1. **Control Performance**: With standard AI recommendations with explanation, participants achieve 75% accuracy
   - This was calculated from human performance in Germann and Merkle's (2019) fund manager data
2. **Treatment Effect**: Alternative explanation delivery methods provide a 3 percentage point improvement (75% → 78%)
   - Assumes that promoting engagement leads to better calibration of AI reliance
   - Large enough to be practically meaningful for investment decisions
3. **Individual Differences**: Between-subject standard deviation of 0.177 percentage points
   - Derived from Germann and Merkle's (2019) fund manager performance data
4. **Learning Effects**: Small practice effect of 1 percentage point per round
   - Participants may improve slightly through experience with the task
   - Controlled by randomising condition order across participants

1.5 Statistical Considerations

We have 5 key comparisons, all versus the control (Explanation) condition. With 5 comparisons, we face increased Type I error risk.

- **Primary analysis**: Individual comparisons at $\alpha = 0.05$
- **Corrected analysis**: Bonferroni correction ($\alpha = 0.01$) for family-wise error control

Target Power: We aim for 90% statistical power: a 90% chance of detecting true effects when they exist.

▶ Show code

▶ Show code

2 Data simulation

The simulation creates synthetic experimental data based on our design assumptions. Each participant makes 36 binary decisions (correct/incorrect) across 6 conditions, with individual differences and small learning effects included.

Output formats:

- Binary: Individual decision outcomes (for GLMM/GEE analysis)
- Scores: Aggregated correct decisions per condition (for paired t-tests)

▶ Show code

Rather than simulate data repeatedly during power calculations, we pre-generate all required datasets once. This reduces computation time and ensures identical data across the three analysis methods.

▶ Show code

▶ Show code

3 Paired t-test on Scores

This method aggregates each participant's decisions into scores per condition (0-6 correct), then compares treatment scores to control scores using paired t-tests. Simple and robust, but loses information by aggregating binary decisions.

▶ Show code

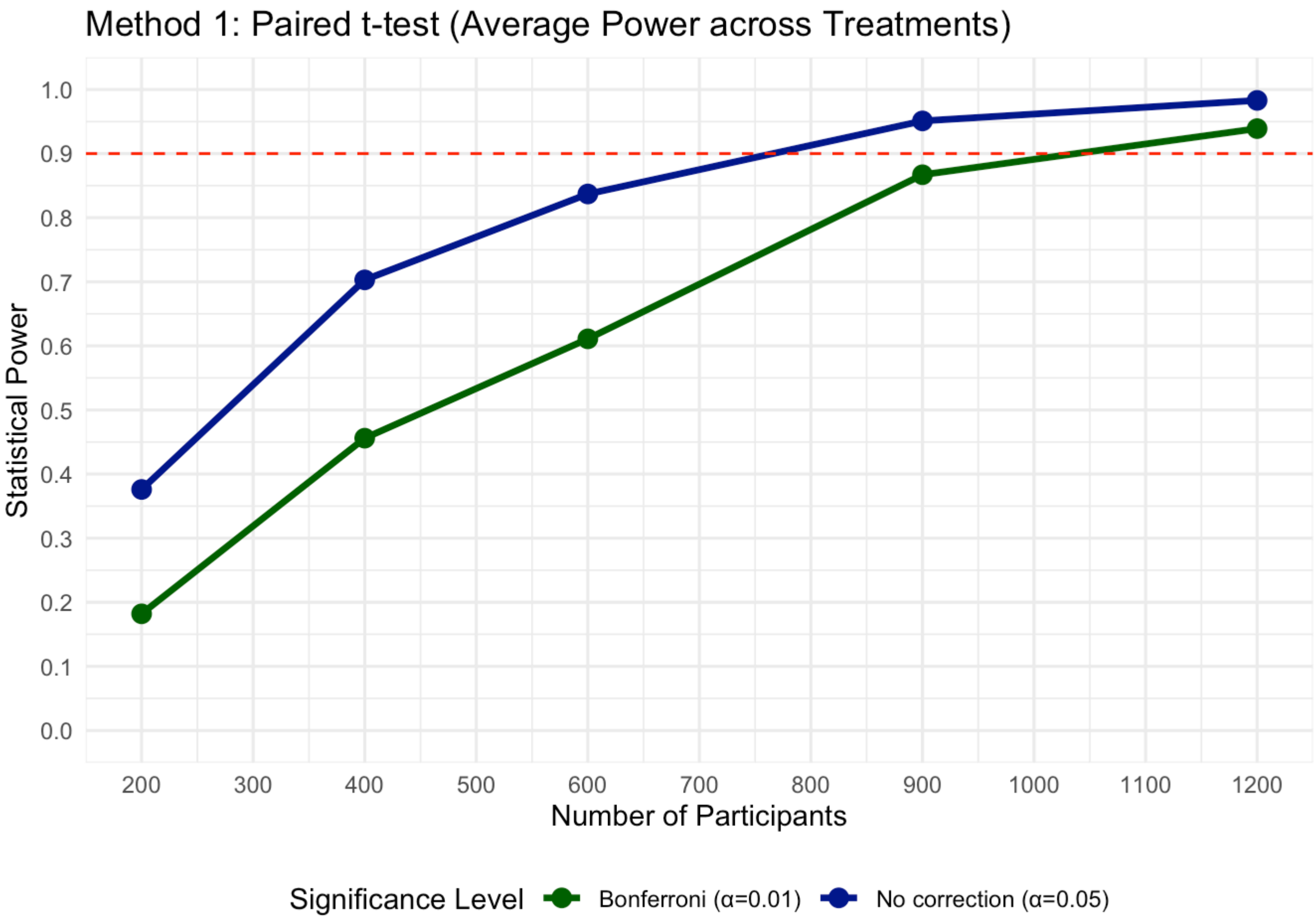
▶ Show code

▶ Show code

Power by Sample Size (Average across 5 treatments)

Sample Size	α=0.05	α=0.01
200	0.376	0.182
400	0.703	0.456
600	0.837	0.611
900	0.951	0.867
1200	0.983	0.939

▶ Show code



4 Mixed Effects Models (GLMM)

This method models individual binary decisions using logistic regression with random participant effects. More statistically efficient than Method 1 but requires distributional assumptions and can have convergence issues.

▶ Show code

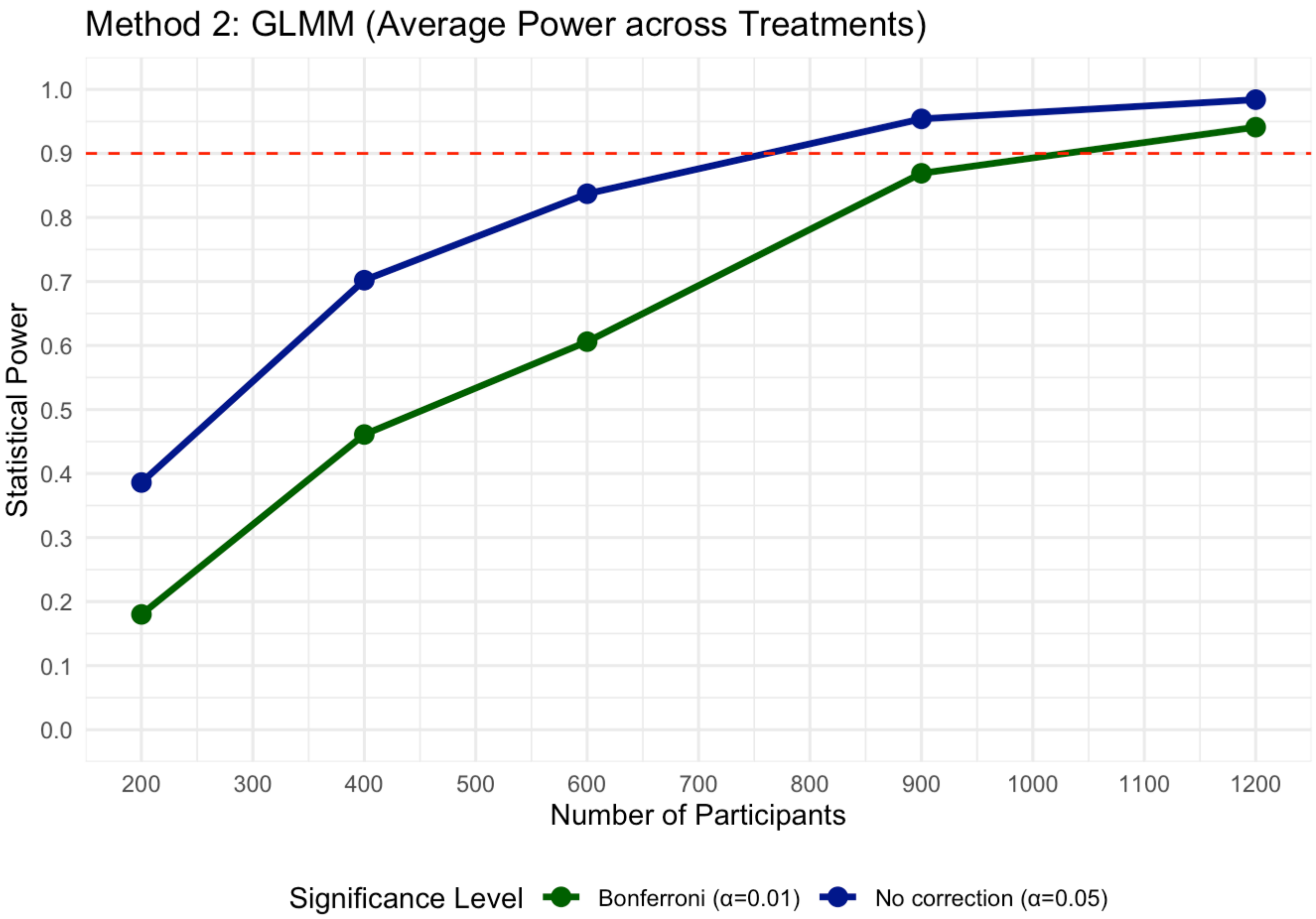
▶ Show code

▶ Show code

Power estimates using GLMM

Sample Size	α=0.05	α=0.01
200	0.386	0.180
400	0.702	0.461
600	0.837	0.606
900	0.954	0.869
1200	0.984	0.941

▶ Show code



4.1 Validation with `simr`

The `simr` package provides an independent validation of our GLMM power estimates. It fits a template model and systematically varies sample size to generate power curves.

First, we need to create a base model using the simulated data.

▶ Show code

We then extend this model to simulate the power curve across a range of sample sizes.

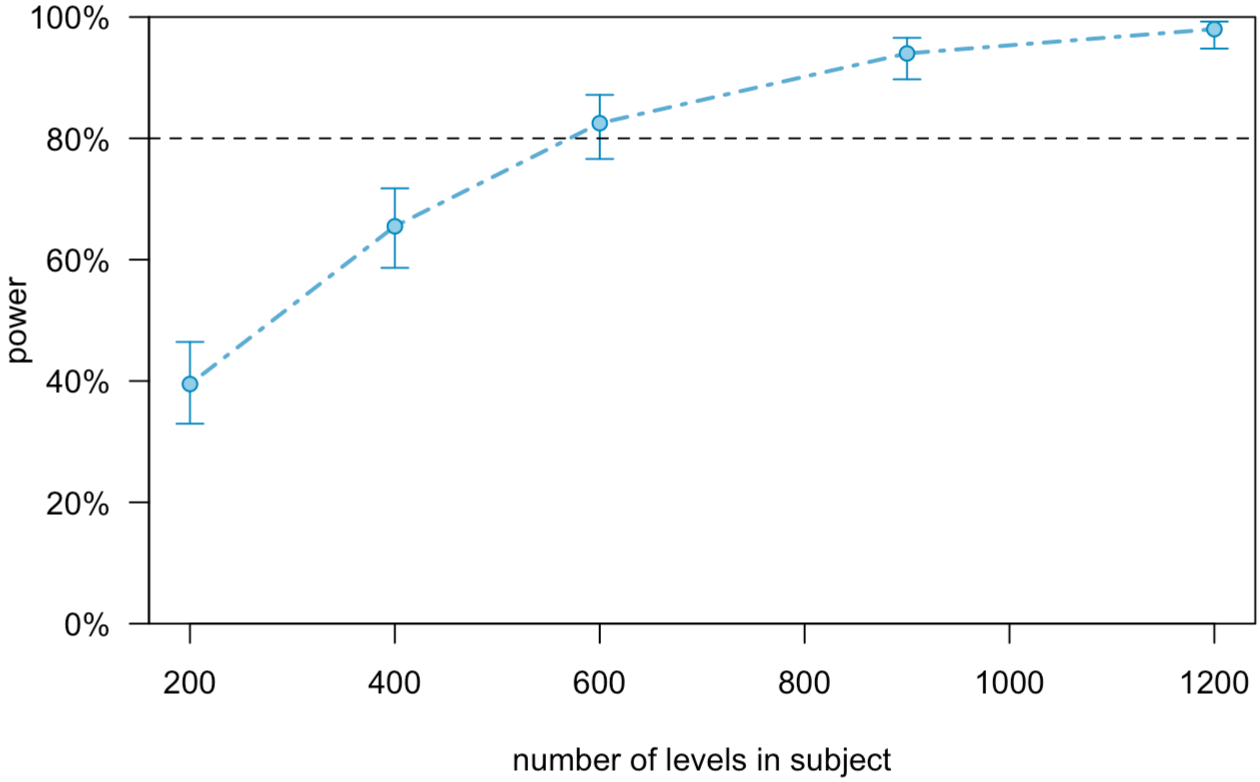
▶ Show code

We then plot the power curve to visualize how the power changes with sample size.

▶ Show code

```
Power curve with standard alpha (0.05):
```

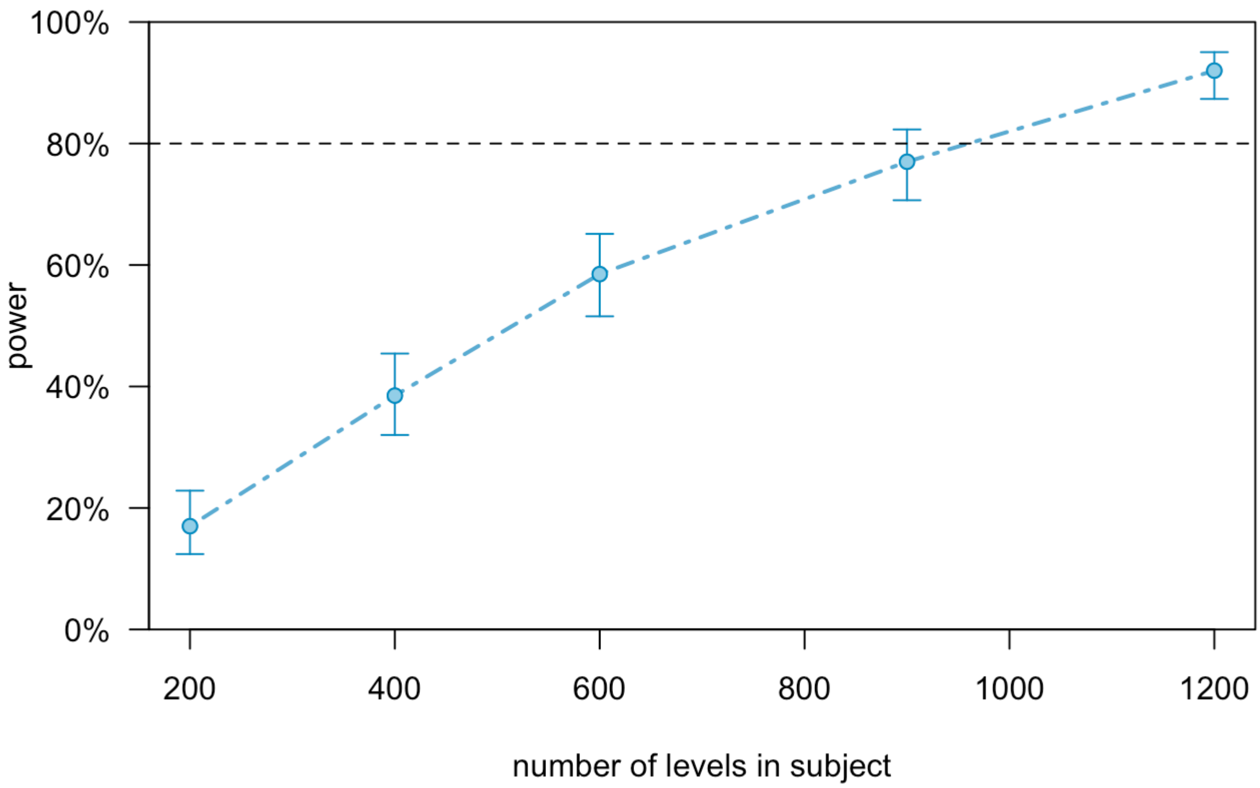
▶ Show code



▶ Show code

```
Power curve with Bonferroni alpha (0.01):
```

▶ Show code



The following code extracts the power values from the `simr` power curve and formats them into a summary table.

▶ Show code

Power using simr at both significance levels

Sample Size	Power (α=0.05)	95% CI	Power (α=0.01)	95% CI
200	39.5%	[32.7% - 46.6%]	17.0%	[12.1% - 22.9%]
400	65.5%	[58.5% - 72.1%]	38.5%	[31.7% - 45.6%]
600	82.5%	[76.5% - 87.5%]	58.5%	[51.3% - 65.4%]
900	94.0%	[89.8% - 96.9%]	77.0%	[70.5% - 82.6%]
1200	98.0%	[95.0% - 99.5%]	92.0%	[87.3% - 95.4%]

5 Generalised Estimating Equations (GEE)

GEE provides a robust alternative to GLMM that estimates population-level effects while accounting for within-subject correlation. More robust to model assumptions than GLMM but potentially less efficient.

▶ Show code

▶ Show code

▶ Show code

Power estimates using GEE

Sample Size	α=0.05	α=0.01
200	0.389	0.183
400	0.703	0.457
600	0.845	0.608
900	0.953	0.868
1200	0.982	0.940

▶ Show code