Power Analysis for CBA experiment

Iñigo

May 26, 2025

1 Methodological Approach to Power Analysis

1.1 Experimental Design Consideration

The experiment is structured such that each participant will experience all six experimental conditions (one control condition and five distinct AI-supported treatment conditions). The order of these conditions will be randomized across participants. For the primary research objective—assessing the effectiveness of each treatment relative to the control by comparing the proportion of *ex-ante* optimal choices—this constitutes a **within-subjects design**.

1.2 Key Parameters for Sample Size Estimation

The sample size calculations are guided by the following conventional parameters:

- **Significance level (α):** An alpha of 0.05 (two-tailed). *** Consider adopting a Bonferroni-corrected $\alpha_{\text{adj}} = 0.01$ per comparison would provide more stringent control over the family-wise error rate and would consequently require a larger sample size ***.
- **Statistical power ($1 - \beta$):** A power of 0.80 is targeted.
- **Effect size conventions:** The effect size is conceptualized through the anticipated improvement (Minimum Detectable Effect - MDE) over a baseline rate of optimal choices (gamma, P_1). This is translated into Cohen's h for assessing the magnitude of the difference between two proportions (P_1 and $P_2 = P_1 + \text{MDE}$).

2 Sample Size Estimates Across Scenarios

The table below details the Cohen's h values and the corresponding estimated number of participants (N) required for various combinations of baseline optimality (γ) and MDEs.

Important Note on Sample Size Interpretation: The sample sizes (N) presented in Table 1 are calculated based on a z-test for two *independent* proportions. As we have a within-subjects design, where measurements are repeated, the required sample size is **smaller** than these estimates, especially if there is a positive correlation between participants' responses across conditions. Therefore, **these figures should be viewed as *conservative upper-bound estimates* for the number of participants needed**.

***** Idk why, but IA (Gemini and Claude) are giving me N numbers way smaller than G*Power manually. I double checked the inputs *****

- γ is the proportion of *ex ante* optimal choices in the Control group (no AI support).
- MDE, minimum detectable effect, is the smallest increase in mean *ex ante* optimality we are looking for in **any one** treatment group, against the control group.

Table 1: Estimated Sample Sizes (N) for Detecting MDEs from Baseline Optimality (γ)

Baseline γ (P_1)	MDE	Treatment P_2	Cohen's h	Participants N (Indep. Groups Test)
0.70	0.050	$0.70 + 0.050 = 0.750$	0.112	≈ 621
0.70	0.075	$0.70 + 0.075 = 0.775$	0.171	≈ 266
0.70	0.100	$0.70 + 0.100 = 0.800$	0.232	≈ 145
0.75	0.050	$0.75 + 0.050 = 0.800$	0.120	≈ 536
0.75	0.075	$0.75 + 0.075 = 0.825$	0.184	≈ 229
0.75	0.100	$0.75 + 0.100 = 0.850$	0.252	≈ 123
0.80	0.050	$0.80 + 0.050 = 0.850$	0.132	≈ 436
0.80	0.075	$0.80 + 0.075 = 0.875$	0.205	≈ 185
0.80	0.100	$0.80 + 0.100 = 0.900$	0.284	≈ 98

*Note: Sample sizes (N) are calculated per group for an independent two-proportion z-test framework, assuming $\alpha = 0.05$ (two-tailed) and power=0.80. For the planned within-subjects design, the required number of participants is anticipated to be smaller. $P_2 = P_1 + MDE$. Calculations based on G*Power outputs for two independent proportions.*

3 Methodology for Within-Subjects Power Analysis

3.1 The McNemar Test for Paired Proportions

For comparing the proportion of optimal choices between two paired conditions (e.g., Control vs. a Treatment) within the same participant, the McNemar test is the appropriate statistical test. This test specifically evaluates changes in proportions for paired nominal data and focuses on the discordant pairs – instances where a participant's choice optimality differs between the two conditions.

Let P_1 be the proportion of optimal choices in the Control condition (baseline "gamma") and P_2 be the proportion of optimal choices in the Treatment condition ($P_2 = P_1 + \text{MDE}$, where MDE is the Minimum Detectable Effect).

The underlying data for a McNemar test can be visualized in a 2x2 table for any pair of compared conditions:

		Treatment Condition	
		Optimal (Success)	Non-Optimal (Failure)
Control Condition	Optimal (S)	p_{SS}	p_{SF}
	Non-Optimal (F)	p_{FS}	p_{FF}

Where:

- p_{SS}: Proportion of participants optimal in *both* Control and Treatment.
- p_{FF}: Proportion of participants non-optimal in *both* Control and Treatment.
- p_{SF}: Proportion optimal in Control but non-optimal in Treatment (Control Success, Treatment Failure).

- p_{FS}: Proportion non-optimal in Control but optimal in Treatment (Control Failure, Treatment Success).

The marginal probabilities are $P_1 = p_{SS} + p_{SF}$ and $P_2 = p_{SS} + p_{FS}$. The McNemar test focuses on the discordant cells p_{SF} and p_{FS}. The difference in marginal proportions is $MDE = P_2 - P_1 = p_{FS} - p_{SF}$.

3.2 Estimating Discordant Proportions for GPower

To use GPower for the McNemar test, we need to provide estimates for p_{SF} and p_{FS}. Knowing P_1, P_2, and thus the MDE, gives one equation ($p_{FS} - p_{SF} = \text{MDE}$) with two unknowns. An additional assumption is required, typically by estimating p_{SS} (the proportion optimal in both conditions). This p_{SS} reflects the agreement or correlation between a participant's responses in the two conditions.

1. Valid range for p_{SS}: The proportion p_{SS} must satisfy:

$$\max(0, P_1 + P_2 - 1) \leq p_{SS} \leq \min(P_1, P_2)$$

2. Calculating p_{SF} **and** p_{FS} **from** p_{SS}: Once a plausible value for p_{SS} is chosen (e.g., based on pilot data, literature, or by exploring a range of values), the discordant proportions are:

- $p_{SF} = P_1 - p_{SS}$
- $p_{FS} = P_2 - p_{SS}$

3. Assumption for Table Values: For the sample size estimates presented in Table **??**, p_{SS} was assumed to be the **mid-point** of its valid mathematical range for each specific (P_1, P_2) scenario. This represents a "medium" or "moderate" level of agreement.

3.3 Example Derivation of Discordant Pairs

Consider $P_1 = 0.750$ and a target $P_2 = 0.800$ (MDE = 0.050).

- Minimum $p_{SS} = \max(0, 0.750 + 0.800 - 1) = 0.550$.
- Maximum $p_{SS} = \min(0.750, 0.800) = 0.750$.
- Assumed mid-range $p_{SS} = (0.550 + 0.750)/2 = 0.650$.
- $p_{SF} = P_1 - p_{SS} = 0.750 - 0.650 = 0.100$.
- $p_{FS} = P_2 - p_{SS} = 0.800 - 0.650 = 0.150$.
- (Check: $p_{FS} - p_{SF} = 0.150 - 0.100 = 0.050 = \text{MDE}$).

These values ($p_{SF} = 0.100, p_{FS} = 0.150$) are then used in GPower.

3.4 Power Analysis Parameters for GPower

- **GPower Test Details:**
  - Test family: 'Exact tests'
  - Statistical test: 'Proportions: Inequality, two dependent groups (McNemar test)'
  - Type of power analysis: 'A priori: Compute required sample size'
- **Input Parameters for GPower:**

– Tail(s): 'One' (hypothesizing improvement, $P_2 > P_1$, thus $p_{FS} > p_{SF}$).
– Proportion p1 (GPower's $P(+-)$): This corresponds to p_{SF}.
– Proportion p2 (GPower's $P(-+)$): This corresponds to p_{FS}.
– Statistical power ($1 - \beta$): Set to 0.80.
– Significance levels (α):
  * For a single comparison: $\alpha = 0.05$ (one-tailed).
  * For multiple comparisons (e.g., 5 treatments vs. control): A Bonferroni-corrected $\alpha_{\text{adj}} = 0.01$ (i.e., 0.05/5) is used to illustrate the impact on sample size.

The sample sizes (N) reported in Table **??** represent the total number of participants (pairs) required for a single paired comparison under the specified alpha level.

4 Sample Size Estimates (McNemar Test with Alpha Adjustment)

Table 2: Estimated N Participants (McNemar Test, One-Sided, 80% Power)

Baseline γ (P_1)	MDE	Treat. P_2	Assumed p_{SS} (Mid)	p_{SF}	p_{FS}	N Participants ($\alpha = 0.05$, 1-sided, 80% Pow.)	N Participants ($\alpha = 0.01$ for 5 comp., 1-sided, 80% Pow.)
0.700	0.050	0.750	0.575	0.125	0.175	≈ 204	≈ 314
0.700	0.075	0.775	0.5875	0.1125	0.1875	≈ 111	≈ 170
0.700	0.100	0.800	0.600	0.100	0.200	≈ 71	≈ 109
0.750	0.050	0.800	0.650	0.100	0.150	≈ 153	≈ 233
0.750	0.075	0.825	0.6625	0.0875	0.1625	≈ 84	≈ 128
0.750	0.100	0.850	0.675	0.075	0.175	≈ 54	≈ 81
0.800	0.050	0.850	0.725	0.075	0.125	≈ 119	≈ 181
0.800	0.075	0.875	0.7375	0.0625	0.1375	≈ 67	≈ 101
0.800	0.100	0.900	0.750	0.050	0.150	≈ 44	≈ 66

Note: N = total number of participants (pairs) for one specific paired comparison. Test is one-sided, Power=0.80. $P_2 = P_1 + MDE$. p_{SS} is assumed to be the mid-point of its valid range for each (P_1, P_2) pair: $[\max(0, P_1 + P_2 - 1), \min(P_1, P_2)]$. *Consequently, $p_{SF} = P_1 - p_{SS}$ and $p_{FS} = P_2 - p_{SS}$. The final column indicates N needed per comparison if using $\alpha = 0.01$ (one-sided) to account for 5 primary comparisons (Bonferroni correction).*

Table 3: Estimated N Participants (McNemar Test, **Two-Sided, 90% Power**)

Baseline γ (P_1)	MDE	Treat. P_2	Assumed p_{SS} (Mid)	p_{SF}	p_{FS}	N Participants ($\alpha = 0.05$, 2-tailed, 90% Pow.)	N Participants ($\alpha = 0.01$ for 5 comp., 2-tailed, 90% Pow.)
0.700	0.050	0.750	0.575	0.125	0.175	≈ 270	≈ 394
0.700	0.075	0.775	0.5875	0.1125	0.1875	≈ 145	≈ 210
0.700	0.100	0.800	0.600	0.100	0.200	≈ 92	≈ 133
0.750	0.050	0.800	0.650	0.100	0.150	≈ 202	≈ 293
0.750	0.075	0.825	0.6625	0.0875	0.1625	≈ 109	≈ 157
0.750	0.100	0.850	0.675	0.075	0.175	≈ 69	≈ 99
0.800	0.050	0.850	0.725	0.075	0.125	≈ 157	≈ 227
0.800	0.075	0.875	0.7375	0.0625	0.1375	≈ 86	≈ 124
0.800	0.100	0.900	0.750	0.050	0.150	≈ 56	≈ 80

Note: N = total number of participants (pairs) for one specific paired comparison. Power=0.90. $P_2 = P_1 + MDE$. p_{SS} is assumed to be the mid-point of its valid range: $[\max(0, P_1 + P_2 - 1), \min(P_1, P_2)]$. *$p_{SF} = P_1 - p_{SS}$; $p_{FS} = P_2 - p_{SS}$. The final column indicates N needed per comparison if using $\alpha = 0.01$ (two-tailed) to account for 5 primary comparisons (Bonferroni correction).*

Table 4: Estimated N Participants (McNemar Test, One-Sided, 90% Power)

Baseline γ (P_1)	**MDE**	**Treat. P_2**	**Assumed p_{SS} (Mid)**	p_{SF}	p_{FS}	**N Participants ($\alpha = 0.05$, 1-sided, 90% Pow.)**	**N Participants ($\alpha = 0.01$ for 5 comp., 1-sided, 90% Pow.)**
0.700	0.050	0.750	0.575	0.125	0.175	≈ 258	≈ 376
0.700	0.075	0.775	0.5875	0.1125	0.1875	≈ 138	≈ 199
0.700	0.100	0.800	0.600	0.100	0.200	≈ 88	≈ 127
0.750	0.050	0.800	0.650	0.100	0.150	≈ 193	≈ 280
0.750	0.075	0.825	0.6625	0.0875	0.1625	≈ 104	≈ 150
0.750	0.100	0.850	0.675	0.075	0.175	≈ 66	≈ 95
0.800	0.050	0.850	0.725	0.075	0.125	≈ 149	≈ 215
0.800	0.075	0.875	0.7375	0.0625	0.1375	≈ 82	≈ 118
0.800	0.100	0.900	0.750	0.050	0.150	≈ 53	≈ 75

Note: N = total number of participants (pairs) for one specific paired comparison. Test is one-sided, Power=0.90. $P_2 = P_1 + MDE$. p_{SS} is assumed to be the mid-point of its valid range for each (P_1, P_2) pair: $[\max(0, P_1 + P_2 - 1), \min(P_1, P_2)]$. *Consequently, $p_{SF} = P_1 - p_{SS}$ and $p_{FS} = P_2 - p_{SS}$. The final column indicates N needed per comparison if using $\alpha = 0.01$ (one-sided) to account for 5 primary comparisons (Bonferroni correction).*

5 Discussion and Final Considerations

The sample sizes estimated using the McNemar test are substantially lower than those that would be derived from independent group comparisons, accurately reflecting the increased statistical power inherent in within-subjects designs.

A critical takeaway is that these estimates are highly sensitive to the assumption made for p_{SS} – the proportion of participants making optimal choices in both the control and treatment conditions being compared. A different choice for p_{SS} will alter the calculated values of p_{SF} and p_{FS} (the discordant pairs), and consequently, the required sample size. It is therefore strongly recommended to conduct a **sensitivity analysis** by varying the assumed p_{SS} across its plausible valid range (from minimum to maximum possible agreement) for your key scenarios of P_1 and MDE. This will provide a spectrum of potential sample size requirements and a more robust understanding for planning.

Furthermore, the final column in the table illustrates the impact of adjusting for multiple comparisons. If testing all 5 treatments against the control with a desire to maintain a family-wise error rate of approximately 0.05, applying a Bonferroni correction (leading to $\alpha_{\text{adj}} = 0.01$ per test) significantly increases the number of participants needed for each of those comparisons to achieve the desired power. The decision on which 'N' to target should consider the specific research hypotheses and the approach to managing multiple comparisons.