Cookie Cats A/B Test Analysis

Impact of moving a mobile game progression gate on retention & engagement

Executive Summary

This report analyzes an A/B test from the mobile puzzle game Cookie Cats. The experiment tested whether moving a level gate from 30 → 40 affected player retention and engagement.

Key finding: Moving the gate to level 40 reduced 7-day retention by approximately 0.8 percentage points. While this result is statistically significant ($p < 0.05$), it falls below the pre-specified Minimum Detectable Effect (MDE) of 1.0 pp. Therefore, the change is not practically meaningful, and rollout is not recommended.

Dataset

Source

column	data type	description
`userid`	`int`	Unique player ID
`version`	`str`	Experiment split (`gate_30` = control, `gate_40` = treatment)
`sum_gamerounds`	`int`	Total rounds per player
`retention_1`	`bool`	Active 1 day after installing the game
`retention_7`	`bool`	Active 7 days after installing the game

The dataset contains approximately 90,000 unique players split across experiment groups.

Experiment Design

Experiment groups: Players split into control (`gate_30`) and treatment (`gate_40`).

Primary metric: Day-7 retention (`retention_7`). Tested with a two-proportion z-test.

Guardrail metrics:

- Day-1 retention (`retention_1`), tested with a two-proportion z-test.
- Engagement (`sum_gamerounds`), tested using Mann–Whitney U and Welch's t-test on log-transformed values.
- Treated as a family and corrected for multiplicity using Holm method.

Decision rule: Recommend rollout only if the primary metric is significant at $alpha = 0.05$ and the observed effect is > 1.0 pp, without any negative effect on guardrails metrics.

Power / MDE: The test targeted 80% power to detect a 1.0 pp change. Observed sample size achieves 0.74 pp MDE, a smaller detectable effect than the designed 1.0 pp MDE at 80% power.

Sample Ratio Mismatch check: Chi-square test ($alpha = 0.001$) on assignment counts.

Results

Chi-square $p = 0.008$ passes SRM check but it is in a cautionary range.

Metric	Test	Control	Treatment	Abs Δ (pp/unit)	p-value
Day-7 retention	Two-proportion z	19.02%	18.20%	-0.82 pp (-4.31%)	0.0016
Day-1 retention	Two-proportion z	44.81%	44.22%	-0.59 pp	0.15
Game rounds	Mann–Whitney U	52.46	51.30	-1.16	0.15
	Welch's t (log)	2.888	2.870	-	0.15

Note:

- Guardrail p-values are Holm-adjusted.
- Cohen's $h = 0.02$ for the primary metric, which is negligible in standardized magnitude (below the conventional threshold of 0.20 for a small effect).

See the table in detail

Interpretation:
The difference in primary metric (`retention_7`) is statistically significant ($p = 0.0016$). The treatment group (`gate_40`) shows a -0.82 pp decrease in day-7 retention compared to control (`gate_30`) with 95% confidence that it falls between [-1.33, -0.31] pp, corresponding to a -4.3% relative drop. While the result is statistically significant, the effect is negative and below the practical threshold of 1.0 pp. Cohen's $h = 0.02$ indicates that the magnitude of this effect is negligible. Even though the negative effect is minimal, rollout is not recommended.

Business Impact

Player impact: For every 100,000 new installs, the change corresponds to ~820 fewer retained players at day 7. The 95% confidence interval suggests the true impact is likely between 312 and 1,328 fewer players.

Revenue impact: Assuming an ARPU of $0.50/month, this indicates approximately $410 monthly revenue loss per 100,000 installs.

Limitations & Next Steps

The dataset does not indicate whether users actually reached the level gate, so some players may not have been exposed. The current estimate is the average effect over all players, and only a fraction q of users actually reach the gate, which may dilute experiment results. Assuming that there is no effect on the unexposed players, the exposed-only effect can be calculated as:
`exposed effect = current effect / q`
95% confidence intervals can be similarly calculated.
As can be seen from the table below, if only 50% of players have reached the gate level in their corresponding groups, the observed effect of retention at day 7 would be -1.64 pp with 95% [-2.66, 0.62].

Exposure rate (q)	Exposed effect (pp)	95% CI low (pp)	95% CI high (pp)
20%	-4.10	-6.65	-1.55
30%	-2.73	-4.43	-1.03
40%	-2.05	-3.33	-0.77
50%	-1.64	-2.66	-0.62
60%	-1.37	-2.22	-0.52
70%	-1.17	-1.90	-0.44
100% (population)	-0.82	-1.33	-0.31

Exposure also affects power sensitivity. MDE among exposed players would be roughly equal to `MDE / q`, meaning that it is harder to detect the effect if q is small.

There are no covariates or behavioral segments are available for adjusted analysis or potential confounders.

Future tests should:

- Track exposure (whether a player actually reached the gate).
- Extend analysis to 30-day retention curves.
- Include covariates (region, platform, spend) for adjusted models.

Conclusion

The experiment demonstrates a statistically significant decrease in retention when moving the gate to level 40, though the decrease is minimal in size. Both statistical and business perspectives suggest not rolling out the change.