Experiment Readout — Onboarding Email Sequence v2

Client: B2B SaaS, self-serve tier (illustrative sample — synthetic data)
Test window: Days 1–9 of a planned 21-day run
Primary metric: 14-day activation rate
Analyst: Mohammed S. Rahman
Date: [date]

Bottom line

Do not ship this variant. Re-run it.

The reported win is an artifact of stopping the test early, and the variant is doubling your unsubscribe rate.

	Reported	After correction
Activation lift	+12.6% relative	Not established
p-value	0.0299	Fails the 0.0056 threshold required for 9 sequential looks
Unsubscribe rate	Not reviewed	**+101% relative, p = 0.00012**

The unsubscribe finding is the one that should concern you regardless of what happens with activation. It is far more statistically solid than the win is.

1. What was tested

A rewritten 5-email onboarding sequence (Variant) against the existing 3-email sequence (Control), randomized at signup, 50/50 split.

The test was designed for a 21-day run at 3,278 users per arm — sized to detect a 20% relative lift at 80% power. It was stopped on day 9 at 4,812 (Control) and 4,790 (Variant) after the dashboard showed significance.

2. The primary result, and why it doesn't hold

Arm	Activated	N	Rate
Control	553	4,812	11.492%
Variant	620	4,790	12.944%

Absolute difference **+1.45pp**, 95% CI **[+0.14pp, +2.76pp]**, z = 2.172, **p = 0.0299**. Taken alone, that clears the conventional 0.05 bar.

It doesn't hold, for one reason: **the test was evaluated every day and stopped the first time it crossed 0.05.** That is not a 5% false-positive procedure.

I simulated this exact stopping rule — 9 daily looks, no true effect, nominal $\alpha = 0.05$, 4,000 trials. **The rule fires 19.7% of the time when nothing is happening.** Roughly one in five null tests run this way produces a "winner."

To hold a genuine 5% error rate across 9 looks, a Pocock boundary requires approximately **p < 0.0056**. The observed 0.0299 is five times too large.

Two supporting points:

- **The effect is small even if real.** Cohen's h = 0.044, well below the 0.2 conventional floor for a small effect. The lower CI bound of +0.14pp is roughly break-even against the cost of writing and maintaining two extra emails.
- **The test could not reliably have detected a realistic effect anyway.** At n ≈ 4,800/arm, power to detect a true 8% relative lift is **28%**; for a 12% lift it is **54%**. The design was only adequate for a 20%+ effect, which is optimistic for an email rewrite.

3. The guardrail failure

This was not in the original analysis plan, which is itself the finding worth fixing.

Arm	Unsubscribed	N	Rate
Control	43	4,812	0.894%
Variant	86	4,790	1.795%

+0.90pp absolute (+101% relative), 95% CI [+0.44pp, +1.36pp], **p = 0.00012**.

This survives any correction you care to apply. Two extra emails are burning your list at twice the prior rate. Even if the activation lift were real, you would be trading a fragile 1.45pp activation gain against a robust 0.90pp permanent loss of addressable audience — and the unsubscribe cost compounds across every future campaign, while the activation gain does not.

4. Recommended next steps

1. **Do not roll out.** Revert to Control.
2. **Re-run with a fixed horizon.** 21 days, no early stopping, one analysis at the end. If you need the option to stop early, use an alpha-spending function (O'Brien–Fleming) set before launch, not a dashboard check.
3. **Test 4 emails, not 5.** The unsubscribe curve suggests volume is the driver. A 4-email variant isolates content quality from send frequency.
4. **Register unsubscribe and spam-complaint rate as guardrails** with a pre-committed rollback threshold. My recommendation: halt on any statistically significant increase, regardless of primary-metric performance.
5. **Size honestly.** For a 10% relative MDE off an 11.5% baseline you need roughly 13,000 users per arm. If that is more traffic than you have, the correct conclusion is that this test isn't worth running — not that you should run it underpowered and hope.

Appendix — method

- **Primary test:** two-proportion z-test, pooled variance, two-sided.
- **Confidence intervals:** unpooled normal-approximation on the difference of proportions.
- **Effect size:** Cohen's h (arcsine transform), appropriate for proportions.
- **Sequential-testing correction:** Monte Carlo simulation of the observed stopping rule (9 equally spaced looks, 4,000 trials) under H_0 with $p = 0.115$, to estimate realized Type I error; Pocock constant boundary for the corrected threshold.
- **Power:** normal-approximation two-proportion power at $\alpha = 0.05$, two-sided.

Assumptions and limits. Randomization is assumed valid at signup — I did not receive assignment logs and could not run a sample-ratio-mismatch check, which should be step one on the re-run. The 14-day activation window means users enrolled in the final days of the test have censored outcomes; this analysis uses only fully observed cohorts. No adjustment for repeat signups.

Reproducible: analysis script and raw counts available on request.