Experiment Readout — Pricing Page Layout, 5-Arm Test

Client: DTC e-commerce, single-SKU subscription (illustrative sample — synthetic data)
Test window: 24 days, complete
Primary metric: Checkout conversion rate (session → completed order)
Analyst: Mohammed S. Rahman
Date: [date]

Bottom line

No variant won. Don't ship D.

But the test surfaced something more valuable than the winner you were hoping for: a large, plausible mobile-specific effect that deserves its own test.

	Reported	After correction
Variant D vs Control	+18.5% relative, p = 0.0438	**Fails** Holm-Bonferroni (threshold 0.0125)
Simultaneous 95% CI	—	**[−0.13pp, +1.30pp]** — includes zero
Effect size	—	Cohen's h = 0.032 (negligible)
Mobile segment, D vs Control	Not analyzed	**+47.9%, p = 0.0011** — exploratory, needs confirmation

1. What was tested

Four pricing-page layouts (A–D) against the current page (Control), randomized by session, 5 arms at approximately 20% traffic each.

Arm	Orders	Sessions	Conversion	vs Control	p
Control	251	8,104	3.097%	—	—
A	255	8,090	3.152%	+1.8%	0.8411
B	245	8,117	3.018%	−2.5%	0.7705
C	265	8,066	3.285%	+6.1%	0.4961
D	298	8,121	**3.669%**	**+18.5%**	**0.0438**

2. Why D is not a winner

You ran four comparisons, not one. With four independent tests at $\alpha = 0.05$, the probability of at least one false positive under a global null is $1 - 0.95^4 \approx$ **18.5%**. Finding one arm at $p = 0.044$ out of four is close to what you'd expect from noise alone.

Applying **Holm-Bonferroni** at a family-wise error rate of 0.05:

Rank	Arm	p	Threshold	Result
1	D	0.0438	0.0125	fail to reject
2	C	0.4961	0.0167	fail to reject
3	B	0.7705	0.0250	fail to reject
4	A	0.8411	0.0500	fail to reject

D fails by a factor of 3.5. I also ran **Benjamini-Hochberg** at FDR = 0.10 — a deliberately more permissive standard, appropriate when you're screening candidates rather than making a ship decision. D fails that too (0.0438 vs a 0.0250 threshold). There is no reasonable correction under which this result stands.

The bootstrap agrees. Resampling the D–Control difference 20,000 times:

- Naive 95% CI: **[+0.02pp, +1.14pp]** — barely excludes zero
- Simultaneous 98.75% CI (Bonferroni-adjusted for 4 arms): **[−0.13pp, +1.30pp]** — includes zero

And the effect is small regardless. Cohen's h = 0.032. Even taking the point estimate at face value, +0.57pp on a 3.1% base is within the range you'd expect from seasonality across a 24-day window.

3. What the test did find

Splitting D against Control by device — **this was not pre-registered, and I am reporting it as a hypothesis, not a result:**

Segment	Control	D	Relative	p	N per arm
Mobile	2.300%	3.402%	+47.9%	0.0011	~4,875
Desktop	4.298%	4.073%	−5.2%	0.6508	~3,238

The pooled result is weak because two real and opposite things are being averaged together. Mobile appears to respond strongly; desktop does not respond at all, or responds slightly negatively.

Why I am not calling this a finding. Post-hoc segmentation is exactly the procedure that generates false positives — if I slice by device, browser, geography, new vs returning, and traffic source, I have run twenty-odd tests and one of them will look impressive. The mobile p-value is small enough that it would survive correction across a handful of segments, which is why it's worth your attention. It is not enough to justify a rollout.

It is also mechanistically plausible, which matters: D's condensed layout removes a comparison table that requires horizontal scrolling on narrow viewports. That's a specific, checkable reason for a mobile-only effect, not a story invented after seeing the number.

4. Recommended next steps

1. **Don't ship D site-wide.** The evidence doesn't support it, and you'd be shipping a possible small negative to desktop, which is your higher-converting segment.
2. **Run a confirmatory mobile-only test: D vs Control.** One comparison, one metric, pre-registered. At 80% power to detect the observed 2.30% → 3.20% effect you need **5,182 sessions per arm**. If you want to be conservative and size for a 20% relative effect instead, that's **18,294 per arm** — worth knowing before you commit, because that's a materially longer run.
3. **Drop A, B, and C.** All three are indistinguishable from Control with tight intervals. They are not underpowered near-misses; they're flat.
4. **Stop running 5-arm tests at this traffic level.** Splitting 40,000 sessions five ways means every arm is underpowered and every comparison needs correction. Two arms at 20,000 each detects roughly what four arms at 8,000 cannot. Sequential head-to-head beats simultaneous multi-arm below ~200k sessions/month.
5. **Pre-register segments.** If mobile/desktop is a real strategic split for you, declare it before launch and size for it. Then it's a result rather than a hypothesis.

Appendix — method

- **Primary test:** two-proportion z-test, pooled variance, two-sided, each variant against Control.
- **Multiplicity:** Holm-Bonferroni step-down at FWER = 0.05 (primary decision rule); Benjamini-Hochberg at FDR = 0.10 reported as a sensitivity check.
- **Bootstrap:** 20,000 resamples of the difference in proportions, percentile method; simultaneous interval at 98.75% = 1 − 0.05/4.
- **Effect size:** Cohen's h.
- **Power/sample size:** normal-approximation two-proportion, $\alpha = 0.05$, two-sided, 80% power.

Assumptions and limits. Session-level rather than user-level randomization means returning visitors may be exposed to multiple arms; this dilutes true effects and is worth fixing before the confirmatory test. No sample-ratio-mismatch check was possible without assignment logs — arm sizes are close enough (8,066–8,121) that gross misallocation is unlikely, but this should be verified. No adjustment for day-of-week or seasonality; the 24-day window covers three full weeks, which mitigates but does not eliminate this.

Reproducible: analysis script and raw counts available on request.