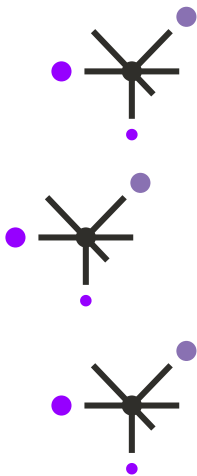


A/B TEST REVIEW

The test said +19%. After rollout, nothing changed.

A three-day review of an onboarding test that won, shipped to everyone, and didn't deliver.

Company	Stillwater, a sleep and meditation app (fictional)
Question	Where did the +19% go? Should we roll back?
Data	Experiment exports, App Store trials, product analytics. All synthetic.
Prepared by	Alex Severianov



This is a sample. The company, people and numbers are invented. The problems, the checks and the reasoning come from real experiment reviews I have done.

00_

Executive summary

Background

Stillwater tested a new “Your sleep plan” onboarding screen for 14 days. The dashboard showed +19% purchases, and the team shipped it to everyone, together with an annual-first paywall from a second test. Four weeks later, there were 680 new paid subscribers instead of the 840 forecast, fewer than before the test. The team was discussing rolling both changes back.

Objective

Explain why the +19% didn’t show up after rollout, find out whether something broke, and recommend whether to roll back.

What was done

I rebuilt the test from raw data: experiment assignments, analytics events, App Store trials and subscriptions, and daily installs by source. I checked the test setup, split the result by install period and by the second test, and compared the four weeks after rollout with the four weeks before the test.

Key findings

- **Nothing broke after rollout.** Trials +5%, paid –4%: almost exactly what the test showed on normal days.
- **The +19% was trial starts, not customers.** Even in the test, paid subscriptions were flat: 181 vs 180 ($p = 0.93$).
- **An outside spike made it look like a win.** A creator video brought 3.4× the usual installs on days 9–12 and 87% of the extra trials. It didn’t repeat after rollout.
- **The paywall forecast used the wrong audience.** Behind the new screen, annual-first adds 3 points, not 7. After rollout: +2 points.
- **The test setup is sound.** Even split, assignment before onboarding, one app version, tracking that matches the App Store.

Conclusion

The rollout did what the test said it would. The test never had a winner: a metric that counted trials and four unusual days made it look like one.

Recommendations

Don’t roll back. Neither change hurts paid conversion, and the paywall still adds about 2 points of annual share. Separate trial and paid events. Keep a 5% holdout for four weeks on every rollout. If onboarding still matters, rerun it on paid conversion, with no paywall test alongside.

01_

What the dashboard promised. What the rollout delivered.

Dashboard	What I found	Verdict
Something broke after rollout	After rollout: trials +5%, paid -4%. In the test, on normal days: +5% and -4%. Traffic and sources unchanged.	Nothing broke
Purchases +19%	The event counts trial starts. Paid subscriptions in the test: 1.40% vs 1.39% of new users (p = 0.93).	Never a win
The new screen works for everyone	87% of the extra trials came from four spike days. On the other days: trials +5%, paid -4%, both within noise.	Audience effect
First-week engagement up	Day-1 return +17%. By day 10: 12.6% vs 12.5%.	Short-lived
Annual-first paywall: +7 pts	+3 pts behind the new screen. After rollout: 41% → 43%.	Smaller, still positive

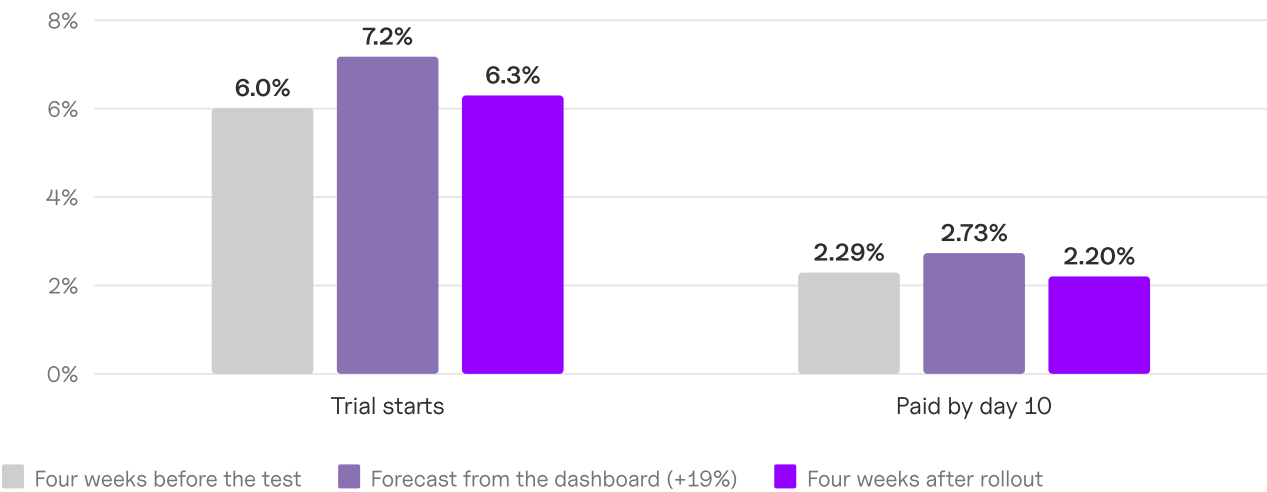
RECOMMENDATION

Don't roll back: it costs work and recovers nothing. Fix the metric, keep a holdout on every rollout, and rerun the onboarding question on paid conversion if it still matters.

02_

After rollout, the numbers landed where the test pointed.

The new screen and the annual-first paywall shipped to everyone on day 15. I compared the four weeks after rollout with the four weeks before the test, with every user followed for ten days.



Four weeks	Before the test	Forecast	After rollout	Change
New users	30,800	–	30,870	0%
Trial starts	1,852	2,210	1,945	+5%
Paid by day 10	705	840	680	–4% p = 0.47
Annual share	41.0%	47.7%	43.2%	+2 pts

The rollout didn’t fail. The forecast was wrong. Trials up about 5%, paid flat, annual share up about 2 points: that is what the test showed once the metric and the spike days are read correctly. The next pages show why.

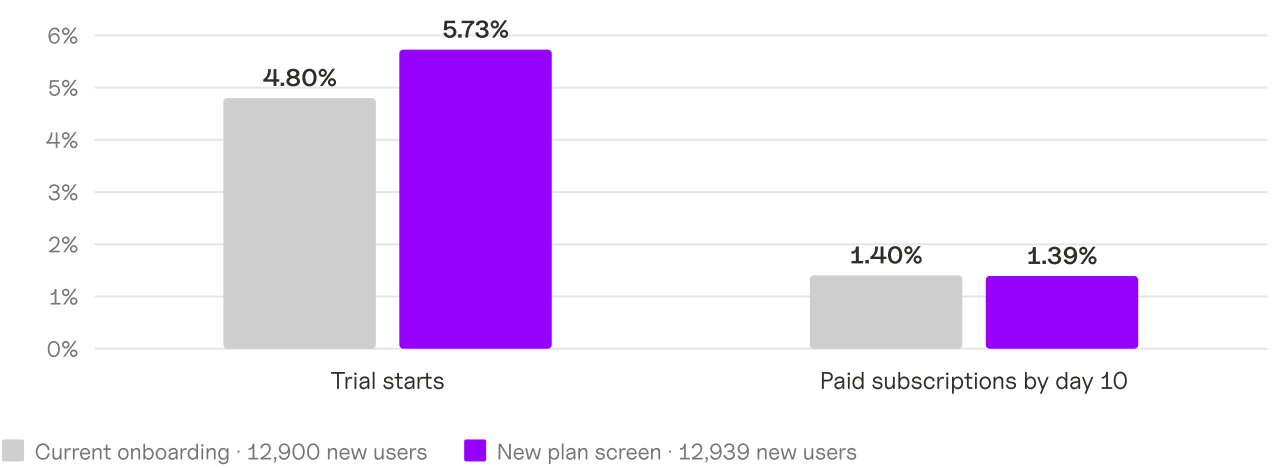
A before-and-after comparison is not randomized. Installs, acquisition sources and season were the same in both periods, and the result matches the test’s own normal days. A holdout would have made this a direct measurement.

03_

The headline metric counted trials, not customers.

The experiment tool reports a “purchase” event. In this app, that event fires when someone starts the free trial. Nobody has paid yet at that point. So the +19% was a lift in trial starts, and the team forecast paying customers from it.

Seven days later, the trial converts or it doesn’t. I followed every test user for ten days and counted actual paid subscriptions.



	Current	New screen	Change	
Trial starts	619	741	+19%	p < 0.001
Paid by day 10	181	180	-1%	p = 0.93
Trial to paid	29.2%	24.3%	-17%	p = 0.04

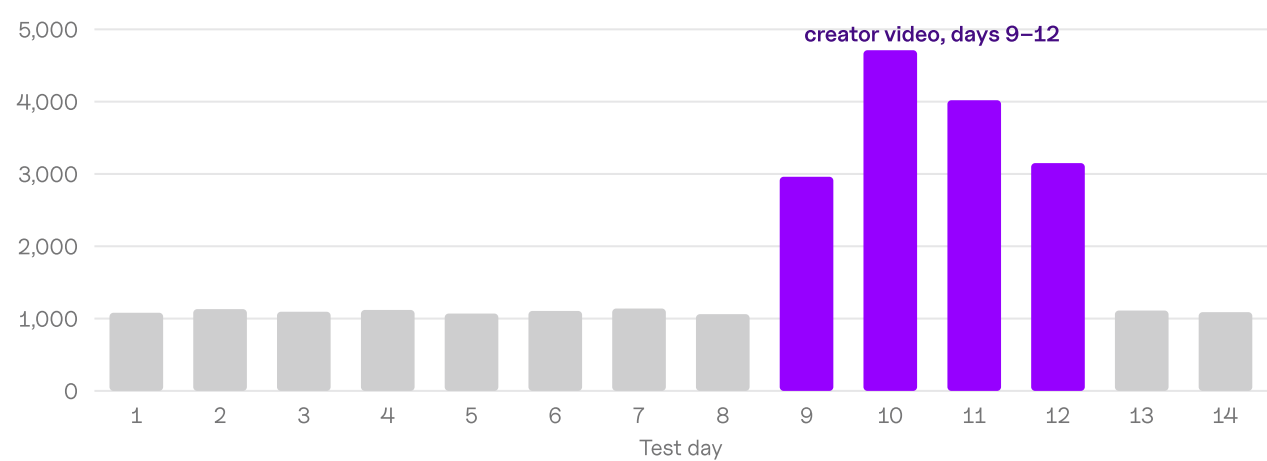
The new screen brought more people into trials, and a smaller share of them paid. The number of paying customers never moved, not in the test and not after rollout.

Rates use all new users in each group as the base, so trial and paid numbers are directly comparable. Every user has at least ten days of follow-up, enough for a seven-day trial to end.

04.

An outside spike changed who was in the test.

Daily installs were steady at about 1,100 a day. On days 9–12 they jumped to 3.4× that. Nobody on the team had planned it. The extra installs came through App Store browse and web referrals, and I traced them to a creator video posted on day 9.



Four days produced 57% of all test users. These people arrived curious, not looking for a sleep app. The new plan screen spoke to them, and many started a trial. Very few paid.

Install days	Group	New users	Trial starts	Paid by day 10
Days 1–8, 13–14	Current	5,512	6.01%	2.29%
	New screen	5,488	6.30% +5%, p = 0.51	2.20% –4%, p = 0.77
Days 9–12, spike	Current	7,388	3.90%	0.74%
	New screen	7,451	5.30% +36%, p < 0.001	0.79% +6%, p = 0.74

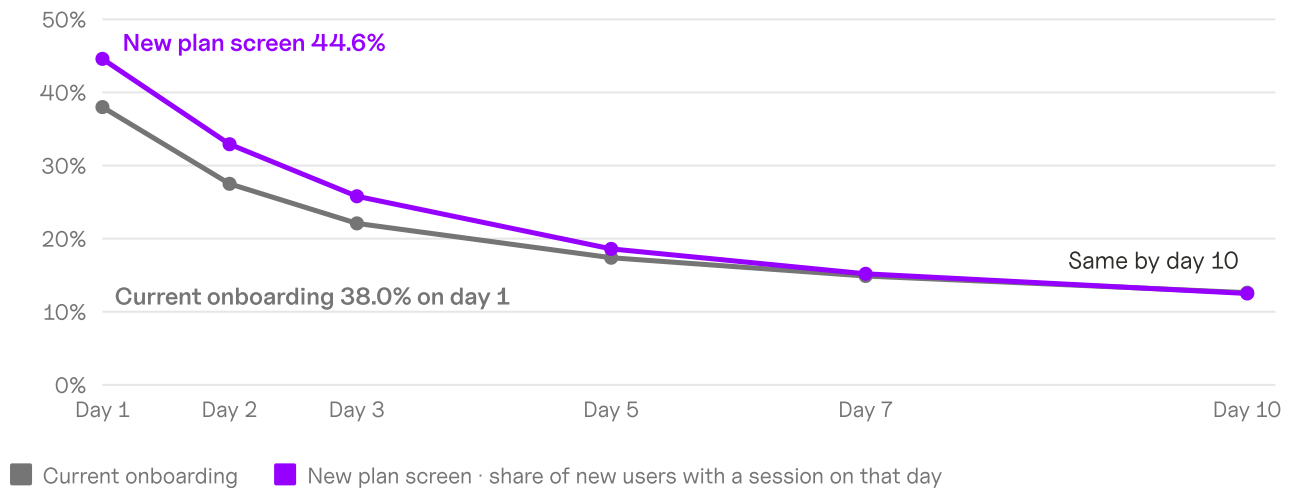
The rollout got normal days, and normal days never showed a win. The spike didn’t repeat: after rollout, installs stayed at about 1,100 a day.

Install day is a proxy for the video audience: spike days also include normal installs. That makes the spike difference look smaller than it really is, not larger. No segment was removed because of its result.

05_

The engagement lift worked like a reminder.

The team also saw more sessions in the first week. That part is true. The new screen ends with a suggested bedtime and a first session for tonight, so people came back the next day.



Day-1 return was 17% higher. The gap shrank every day after that, and by day 10 the groups were the same: 12.6% and 12.5%.

A reminder brings people back once. A habit keeps them. This screen did the first. That still has value: a better first night is a reason to keep the screen. But it was never a retention win, and it doesn't add paying customers.

Return here means any session on that calendar day after install. Paid conversion and day-10 return use the same users. The first-week session count the team saw includes the extra day-1 and day-2 visits, which explains most of its lift.

06_

The paywall forecast used a different audience.

While this test ran, a paywall test compared two layouts: monthly plan first or annual plan first. Its dashboard showed 7 points more people choosing annual, and that went into the forecast. Users were in both tests at once, so I split the paywall result by onboarding group.

Onboarding group	Monthly first: annual share	Annual first: annual share	Effect
Current onboarding	41.0% 127/310	52.1% 161/309	+11 pts
New plan screen	40.2% 149/371	43.2% 160/370	+3 pts
Both, as on the dashboard	40.5%	47.3%	+7 pts

Behind the current onboarding, the annual-first layout worked well. Behind the new screen, it barely moved anything. The likely reason is who reaches the paywall: the new screen pushed more low-intent users into trials, and they picked monthly whatever the layout showed.

The groups are small, and the difference between the two effects is not statistically certain ($p = 0.13$). But both changes shipped together, so the new-screen row is the one that applied. **After rollout, annual share went from 41.0% to 43.2%: the smaller number, not the forecast.**

Annual share is measured among people who started a trial and saw the paywall. Both tests were randomized independently, so the split itself is fair. The problem is reading one test's average without asking who the other test sent to it.

07_

What checked out.

A review is only useful if it also says what is fine. Before looking for explanations, I checked whether the test and the rollout could be trusted.

- ✓ **Even split.** 12,900 vs 12,939 new users. A sample-ratio check shows no problem ($p = 0.81$).
- ✓ **Assignment before onboarding, one app version.** Users were assigned on first launch, and both groups ran the same release.
- ✓ **Tracking matches the source.** Trial events in analytics were within 1% of App Store trial counts.
- ✓ **The rollout shipped correctly.** All users on the new screen and the annual-first layout from day 15. No crashes or errors added.
- ! **Stopped early.** The test was planned for 21 days and stopped on day 14, the first day the dashboard showed significance. It makes a false win more likely.

08_

What I would do next.

Now	Don't roll back. Keep the bedtime suggestion and the annual-first layout.
Measurement	Rename the event to <i>trial_start</i> and add a paid event. Keep a 5% holdout for four weeks on every rollout.
Re-test	If onboarding still matters: three weeks, paid by day 10, acquisition source as a planned split, no paywall test alongside. That detects a 24% change in paid conversion; six weeks gets to 17%.
Spikes	Set an alert on daily installs. When traffic jumps, mark the days and analyze them separately.

09_

How this review was done.

Data used

Experiment assignment export, product analytics events, App Store trial and subscription records, daily install reports by source, for the test and the four weeks before and after it. Joined on user ID, aggregated before analysis.

Definitions

Trial start: a free trial began. **Paid:** a trial converted to a paid subscription by day 10 after install. **Return:** any session on that calendar day. **Annual share:** trial starters choosing the annual plan.

Statistics

Two-sided tests on proportions, 5% threshold. P-values are reported as they are, including the ones that aren't significant. Re-test sizes assume 80% power and a 50/50 split of organic traffic.

Scope and limits

This review covers the test result and the rollout. The onboarding funnel was not reviewed step by step for each group; a full review would add that. Install day is a proxy for the video audience. Ten-day paid conversion says nothing about renewals.

Time

Three days: one to rebuild the test and the rollout from raw data, one to find why they differ from the dashboard, and one to check the explanations and write this up.

What you get

A short message with the answer, a report like this one with every number traceable, and a call to go through it with the team.



Shipped a winner that didn't deliver?

Send it to me. I'll tell you what I would check first.

x91tech.com · Alex Severianov