When Not to Trust a Published A/B Test – an Example

Ronny Kohavi
July 29, 2023

TL;DR: surprising A/B tests that are published should be viewed with skepticism. Here I review a recent example, where my Twyman's Law meter spiked, so I am sharing my observations to raise awareness of trust-related evaluations that others should do when reading or listening to a result of an A/B test.

Introduction

Surprising A/B tests, where a small change resulted in a big change to key metrics, are the hallmark of books, conferences, courses, and articles about A/B testing. My book (https://experimentguide.com) and online class (https://bit.ly/ABClassRKLI) are no exception, as both start with such examples.

The question one should ask is the level of trust you should assign to these. Over the years, I developed healthy skepticism towards extreme results. When these were presented at Microsoft and Airbnb, where I worked, I called Twyman's Law (Any figure that looks interesting or different is usually wrong) many times, and probably 9 out of 10 times we found an issue that invalidated the result. The surprising results I have shared were properly powered and the extreme ones (like the opening example in the book) replicated several times. I previously called to raise the bar on shared A/B tests, so this is a continuation of that effort.

The example I share below comes from GuessTheTest, a site that publishes A/B tests regularly (about every other week). As expected from the need to publish regularly, the quality and trust-level vary. Over time, Deborah O'Malley has improved the evaluation and now regularly highlights trust issues with every experiment. She even posted criticism of extreme results herself. I like the overall site and I am a "Pro Member" because it provides positive value. I do hope that my criticism of this experiment is constructive and helps readers better evaluate the trust level of experiments that are published, and experiments that they run in their own organization.

The A/B Test

The test was run by Optimizely, an A/B Testing Vendor, on their own site. The site uses "Get Started" as a Call to Action (CTA) on its pages in the upper-right, as shown below in Figure 1. Additional details are on GuessTheTest - Which CTA copy won?

The Treatment replaced that copy with "Watch a demo" on the Orchestrate product page, as shown below in Figure 2.

The test ran for 44 days with a 50%/50% design. 22,208 visitors saw the Control and 22,129 visitors saw the Treatment.

The Overall Evaluation Criterion (OEC) was clicks on the button, that is, click-through rate.

The results showed that Control had 0.91% click-through rate and the Treatment had 1.59% click-through rate, a 75% lift.

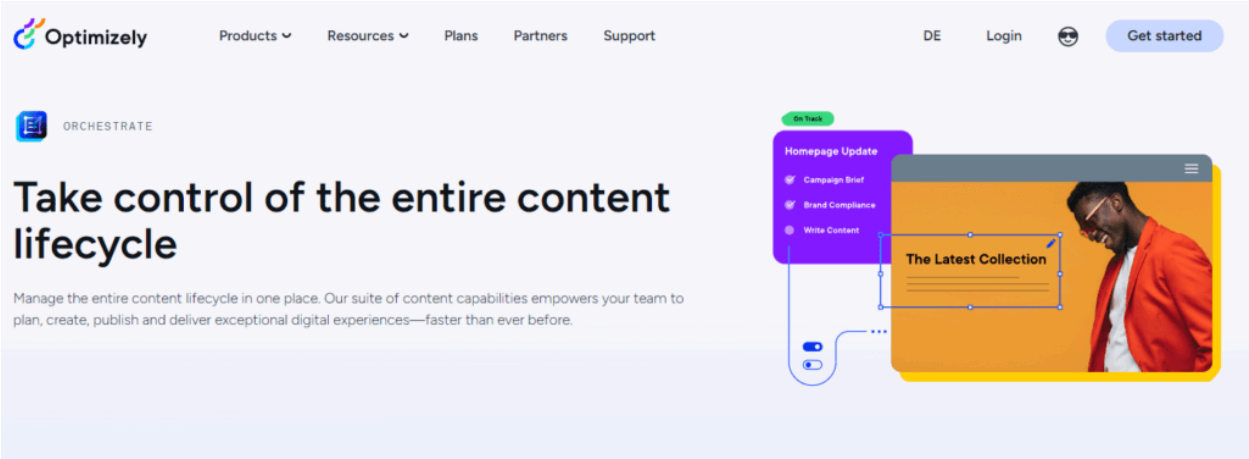


Figure 1: Control with "Get started" in the upper-right

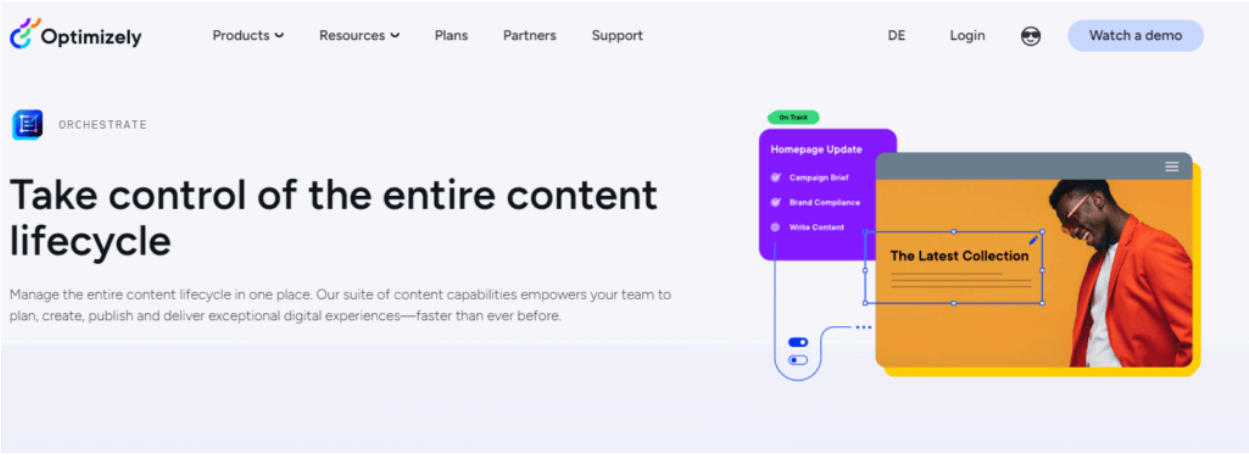


Figure 2: Treatment with "Watch a demo" in the upper-right

The Good

There are several things that were done well.

1. There was only one main factor that changed: the copy.
   The background or font were not changed. The button is slightly bigger for the Treatment, which does tend to increase click-throughs, but it's relatively minor. One could make the two buttons the same size to minimize the possible effect here.
2. The design was maximally powered at 50%/50%.
3. There is no SRM (Sample Ratio Mismatch).
4. The duration is more than a week. I normally recommend a multiple of weeks, and would have recommended 42 days or 49 days, but with a relatively long running time, this unlikely made a material difference relative to other concerns.

The Concerns

1. The biggest concern I have is statistical power. As Deborah herself notes "Sample size calculations should always be done ahead of running the study." She even points to her article on the winner's curse.
   Our article highlighting this point with many references is A/B Testing Intuition Busters.

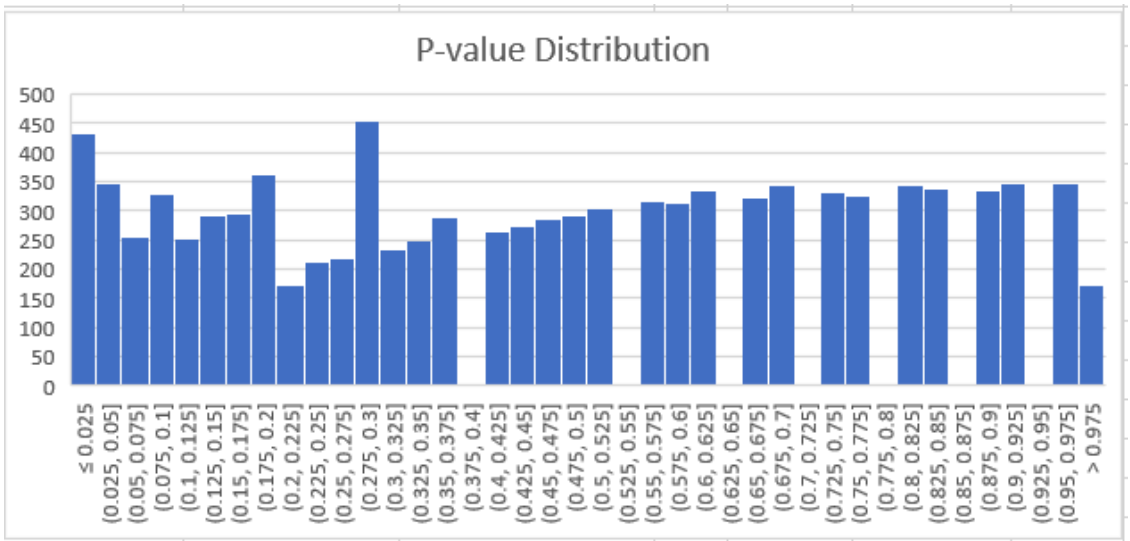
My recommended default as the MDE to plug-in for most e-commerce sites is 5%.
If you're making a big change or have reason to believe (e.g., based on prior patterns) that the change will be big, you might plug in 10%, but a change to CTA doesn't seem to qualify.
Looking at GoodUI for CTA, a site that shows patterns, we see similar patterns in the range of 3.1% to 6.2%:

a. Pattern #13: centered forms & buttons with 5.1% lift
b. Pattern #18: single or alternative buttons with 6.2% lift
c. Pattern #114: less or more visible prices with 3.1% lift

Plugging in the 0.91% click-through rate with a relative MDE of 5% into a power calculator that is mentioned in the article, and it yields a minimum sample size of 688,000 per variation.
The test was run with about 22,000 users, so the experiment is highly under-powered.

The GuessTheTest article computes post-hoc power using 74.7% as the MDE. This is a noisy and misleading as shown in A/B Testing Intuition Busters (Section 5). There is a great article: The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis by Hoenig and Heisey 2001 (official, PDF), which explains this point in detail. Basically, if you have a statistically significant result, the post-hoc power calculation will always show that you have at least 50% power. Since the p-value for this experiment was 0.02, it *appears* that there is enough power to detect a 74.7% change, but it's a catch-22 argument that is incorrect. I strongly disagree with Deborah's use of power-power calculations and her conclusion that the sample size is sufficient.

In this case, the estimated power (spreadsheet) isn't 50% or anywhere closer to 80%, it's 7.3% !
What happens when you have an under-powered experiment? The winner's curse: if you do get a statistically significant result, the treatment effect estimate is likely to be highly exaggerated.
Here is the distribution of p-values you will get from running 10,000 experiments with 5% lift with the above 7.3% power (spreadsheet):



Stat-sig results	13.7%		Average abs(Lift)	22.9%
Max stat-sig Lift	55.5%		Min stat-sig lift	14.3%
Wrong sign when stat-sig	4.8%		Average exaggeration	4.6

Figure 3: P-value distribution with 7.3% power

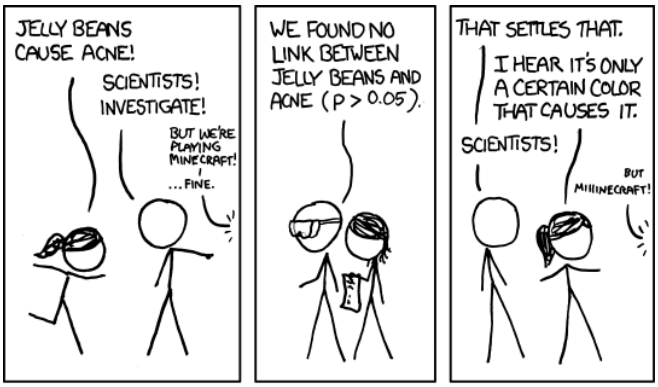
The distribution is close to uniform, which is the p-value distribution when there is no effect (when the Null is true). If you do get a statistically significant result, the average lift you will get is 23% (instead of 5%), which exaggerates the lift by a factor of 4.6.

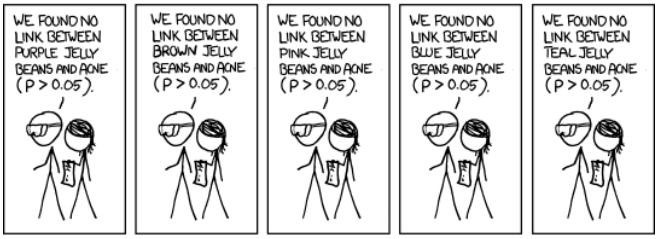
There is a high probability that this is simply a false positive result.

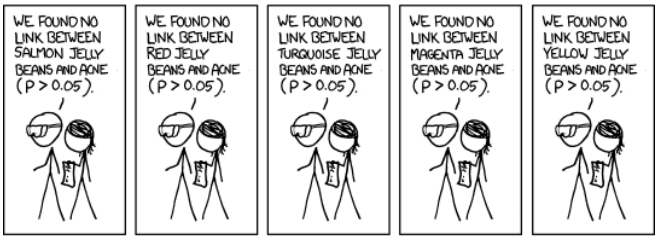
2. The story of the test seems to imply that multiple pages were evaluated. Quoting from the article:

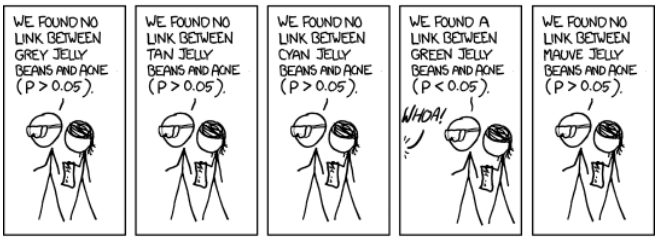
> …the wording "Get Started" worked best globally, across all pages on the site. However, the team wondered if this same finding would hold true on specific product pages.

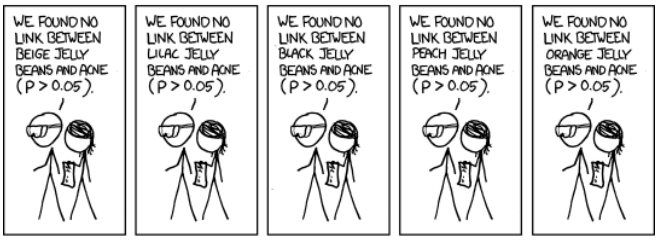
This is the famous Jelly-Bean example by XKCD, where testing 20 colors of Jelly-beans will likely lead to a stat-sig result at the 95% level:

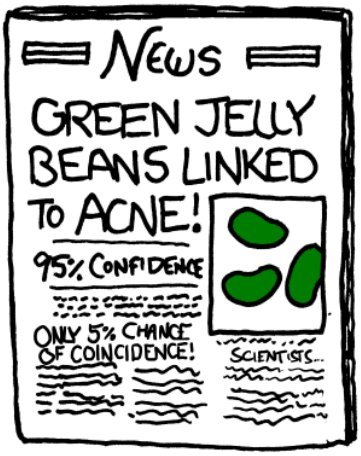

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

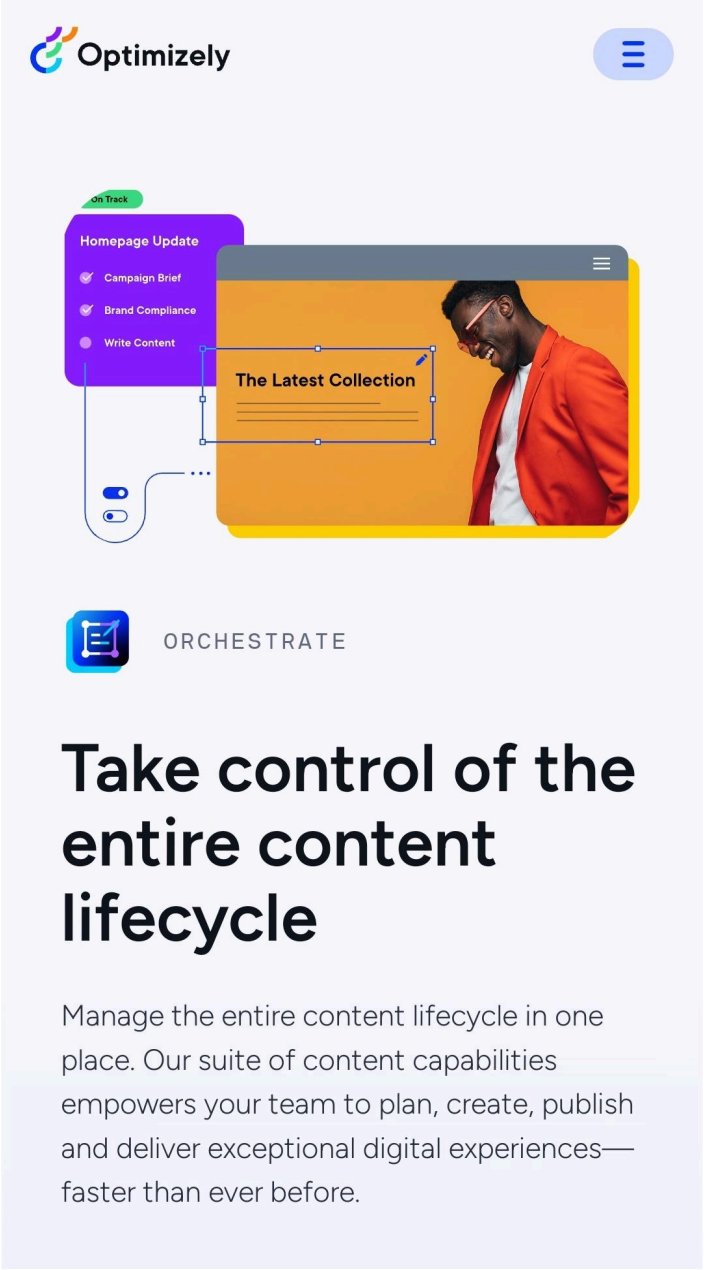
While the p-value claimed in the article is 0.02, we don't know how many pages were evaluated and the text seems to imply that this wasn't the only one.

When testing multiple-hypotheses, a lower alpha threshold should be used. For example, if 10 independent pages were tested, Bonferroni correction recommends using an alpha threshold of 0.005 to determine statistically significant.

Alex Deng suggested that to raise the trust level, we should know how many other product pages were evaluated and the distribution of lifts in these other versions.

3. Visiting the page on Optimizely's site: https://www.optimizely.com/products/orchestrate/, the control is showing. If Optimizely believed the 75% lift number from the Treatment, they should have launched this quickly, so this is a red flag.

4. The button does not seem to show up on smaller viewpoints, a great point made by Lukas Vermeer.
On my 6.8 inch Samsung 23 Ultra phone, this is what I see



The button is shown below the fold.
A similar design, where the button is at the bottom is shown on a narrow PC browser windows.
If the user doesn't see the copy, any treatment effect is likely to be diluted, as the treatment effect for those users is zero.
It isn't clear from the description whether this was a PC-only test or if triggering was employed to limit to users who actually saw the button.

Summary

The article claims that “Overall, this test appears to satisfy the criteria for a trustworthy test result and is an exemplary study.” I disagree and I provided three reasons for my skepticism.

The title of Section 3 of our paper A/B Testing Intuition Busters is “Surprising Results Require Strong Evidence – Lower p-values.” I’ve seen tens of thousands of experiments at Microsoft, Airbnb, and Amazon, and it is extremely rare to see any lift over 10% to a key metric. The metric here is local (click on a button), so larger lifts are possible, but the 75% lift in this example seems unexpectedly large. With the concerns above, I’m skeptical this is a trustworthy result. At a minimum, I would recommend a replication run or two, and I suspect this is simply a nice example of the winner’s curse.

Acknowledgments

Thanks to Alex Deng and Lukas Vermeer for early feedback.