https://bit.ly/ABThrowdown

A/B Testing Throwdown: Can You Predict the Outcome of Actual Live Experiments?

Ron Kohavi, Jakub Linowski, Lukas Vermeer
Updated 1/1/2025

Are there predictable treatment effects for some patterns?

Some people say you cannot know if something will work until you test it. We believe this is not entirely true. Some patterns have predictable outcomes for people with enough experience. We want to validate that hypothesis by testing whether the experimentation community can predict the outcomes, starting with two patterns. Your mission, should you choose to accept it, is to pre-register (predict) the treatment effects.

A few months ago, we started the A/B Patterns Reproducibility project (https://bit.ly/trustworthyABPatterns), calling for public replications of five patterns, and volunteering our time to help design and execute the experiments.

We now have two live experiments on two patterns, and we want to challenge the community to predict the treatment effect, study the prediction distribution, and to provide feedback to the community.

To highlight the community goals and encourage some time investment by the participants, for each of the two experiments below, if the treatment effect is statistically significant (p-value <0.05), then the person whose prediction is the closest will get to name a charity related to experimentation that will be awarded $1,000 (see Appendix A).

The two websites on which each experiment is running are part of Coop, Norway's 15th biggest company, with the descriptions provided by the experiment owner:

- Obs (https://www.obs.no), Norway's largest hypermarket, with 31 physical stores throughout the country, sells groceries and non-food; the webshop sells non-food items. The product range varies from PlayStation consoles to gardening tools, sporting equipment, and LEGO®. More or less everything you would ever need.
- Obs BYGG (https://www.obsbygg.no), one of Norway's largest B2C DIY (do-it-yourself) chains with 59 stores throughout the country, sells everything you need for your home improvement project, your garden, or your garage. The product range varies from a broad selection of tools, paint, flooring, windows, doors, and equipment and furniture for your garden.

The two patterns being tested are described briefly below. More details are available in Appendix B and C, including how to see the live variants:

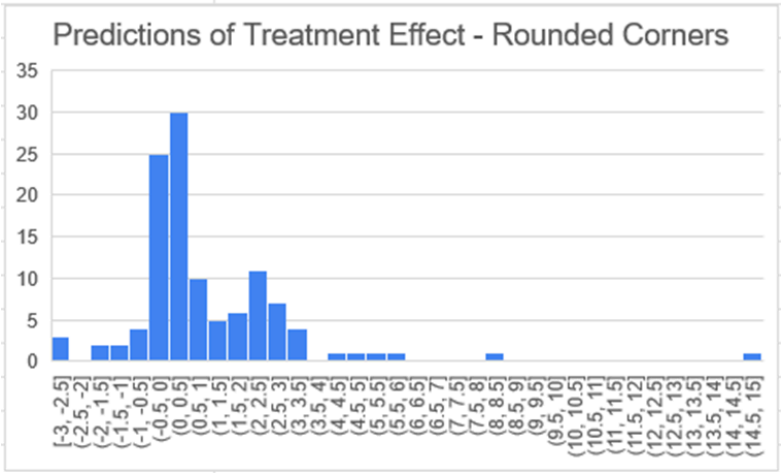
https://bit.ly/ABThrowdown

1. Rounded/square corners. A paper published in a peer-reviewed journal claimed a result that we suspect is highly exaggerated, making it a classic for polling the community and replicating. Details at https://bit.ly/roundedOrSquarePost .

   The OEC (Overall Evaluation Criterion) is the per user conversion from the product page to add-to-cart (Boolean per user). The buttons are either square or rounded.

   ~~To enter your prediction, visit https://bit.ly/ABThrowdown1 [closed 1/1/2025]~~

   Here is the graph of the 115 predictions



   Twenty people submitted exactly zero (impossible to tell from histogram).
   Thirty people submitted a prediction greater than zero, but less than or equal to 0.5%

   The meta-analysis draft analysis is at https://bit.ly/metaAnalysisLiftRK -> Throwdown1 tab (use private/incognito mode if it doesn't seem to work).

   There were over 950,000 users for each variant in the Obs site.
   There were over 1,100,000 users for each variant in the Obs BYGG site.

   The lift is 0.56% with p-value of 0.096, so not statistically significant.
   - For the Obs site, Kameleoon reports a CUPED lift of +0.80% with reliability 99.22% (p-value 0.0078).
   - For the Obs BYGG site, Kameleoon reports a CUPED lift of +0.18% with reliability 36.49 (p-value of 0.635)

   For the second treatment with slightly rounded corners, the meta-analysis lift is 0.65% with p-value 0.054.
   - For the Obs site, Kameleoon reports a CUPED lift of +0.75% with reliability 98.78% (p-value of 0.0122)

- For the Obs BYGG site, Kameleoon reports a CUPED lift of +0.35% with reliability 63.94% (p-value of 0.36)

If both treatments are combined into a larger treatment, that is, assuming their treatment effect is similar, then the meta-analysis shows we have a lift of 0.61% and p-value of 0.038, which is statistically significant (but still smaller by a factor of almost 100 from the journal paper).

Compared to the journal paper at https://doi.org/10.1093/jcr/ucad078, these have much higher statistical power and therefore much more trustworthy. The paper's only reliable randomized controlled experiment, or A/B test, had 474 and 445 visits, so each of our two experiments has about 2,000 *times* more users. They reported a 55% lift to clicks; our estimates are that the treatment effect is about 0.56% to 0.65%, about a factor of 100 smaller. Exaggeration is a common problem with underpowered experiments (https://bit.ly/ABTestingIntuitionBusters Section 4 and Gelman and Carlin's paper at https://www.stat.columbia.edu/~gelman/research/published/retropower_final.pdf).
Note: the other controlled experiment reported in the paper has a massive sample ratio mismatch because Google's algorithm is optimizing the budgets and that assignment is not randomized, hence it is not a trustworthy A/B test.

2. The second pattern is the elimination of the coupon code field.
   The company currently has a few coupons available, but they are rarely used except in specific events. During such a four-day event in December, the experiment will be suspended, but otherwise, the field will be removed.

   The OEC is conversion rate from initiating checkout to purchase (Boolean per user), again a weighted average of the two sites. The experiment triggers only for users who start checkout and can see the coupon code.

   A description of the motivation for this pattern is in https://experimentguide.com Chapter 2, and in Session 2 of the online course: https://bit.ly/ABClassRKLI. A lift of 2.8% is mentioned in these.

   To enter your prediction, visit https://bit.ly/ABThrowdown2

   Happy predicting!

Appendix A – Determining the Winning Predictions and Charities

Here are details of how we intend to name the winners and charities:

1. Each person, identified by LinkedIn URL, can predict one relative treatment effect (e.g., X.X%). If multiple people make the same closest prediction, multiple charities will be named, and the amount will be split approximately equally.
2. The charity chosen by the winner(s) can be any of the following:
    a. Center for Open Science: https://www.cos.io/support-cos (research)
    b. J-PAL: https://www.povertyactionlab.org/support-us (part of MIT)
    c. Test & Learn Community: https://testandlearn.community/donate (research)
    d. A charity related to experimentation, approved by Ron Kohavi. It must be recognized as a 501(c)(3) in the US in Fidelity Charitable. See prior discussions.
3. Kohavi, Linowski, and Vermeer will publish an initial summary after the experiments end, solicit community input for any questions, and then publish the "official" treatment effect for the throwdown. If the experiment is deemed not trustworthy, or the result is not statistically significant, there will be no winner(s) named.
4. The vendor used is Kameleoon, and our plan is to turn on CUPED for the analysis.
5. A person can update their vote multiple times. The last one will count.
6. No one associated with the experiment can participate (e.g., from the company experimenting, from vendor used)

https://bit.ly/ABThrowdown

Appendix B - Rounded/square Corners Experiment Details

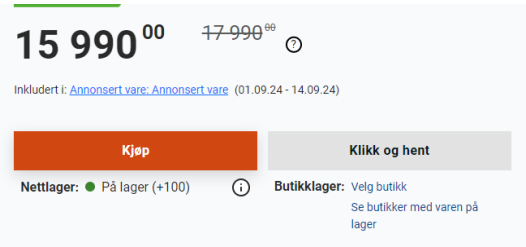
The experiment is now running on the two sites mentioned above (each on web and mobile) in Norway.

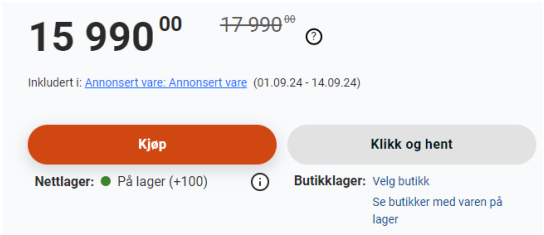
The OEC (Overall Evaluation Criterion) is the per user conversion from the product page to add-to-cart (Boolean per user). Users are triggered when they are on the product page and see the buttons.

There are actually three variants running (a third “slightly round” variation is a treatment that we will eventually report on, but this variation is ignored for the purposes of this throwdown). The experiment is run with 80% power to detect a 2% effect, an MDE (minimum detectable effect) much lower than the 17% to 55% mentioned in the paper.

Because the experiment is running on two websites, we will do a weighted average of the two lifts, as described in the Meta-Analysis book (see post, thanks for Tyler Buffington for correcting my prior phrasing here).

The two variants of interest show different buttons are as follows:

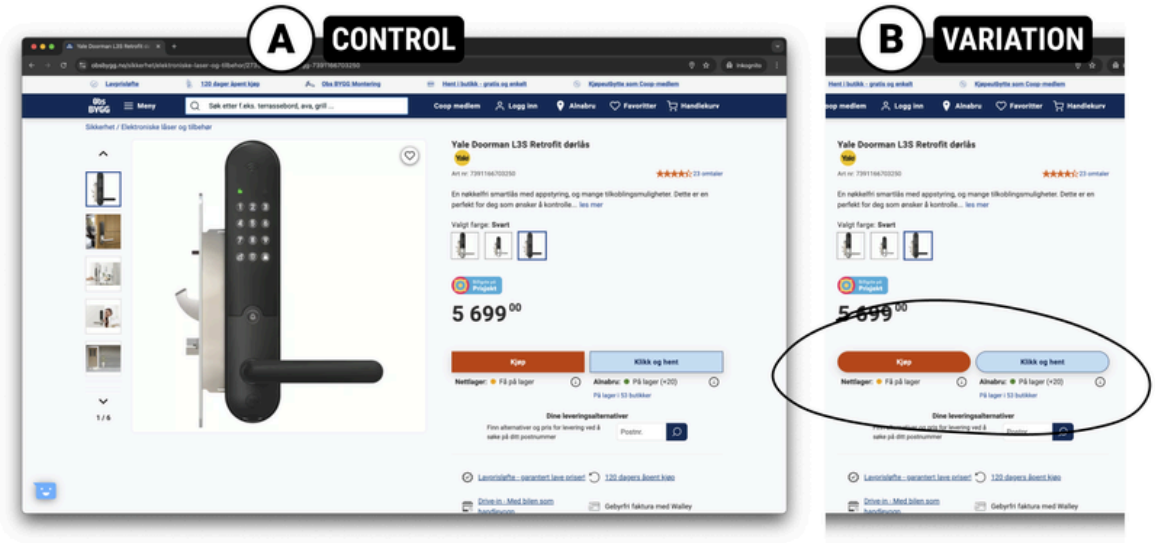




The two buttons indicate either buy, which adds to cart, or Click-and-collect, where you can pick it up in a store. Either click counts as a conversion for the purpose of this experiment.

Zoomed out, it looks like this:

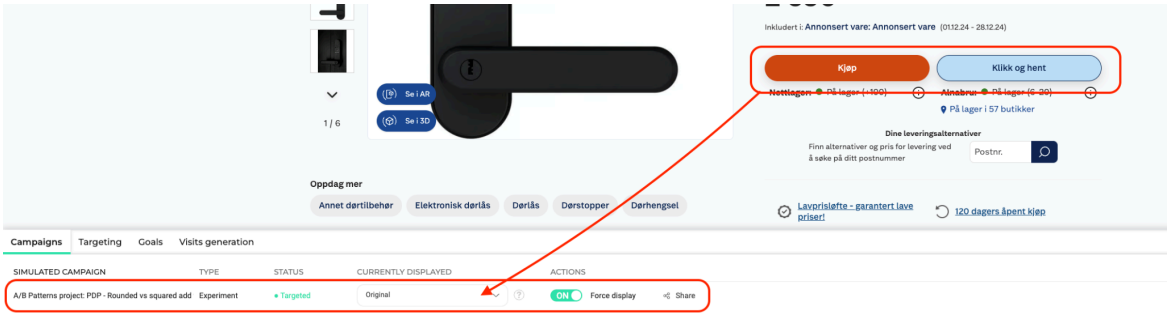
https://bit.ly/ABThrowdown



https://bit.ly/ABThrowdown

To try the variants live, follow the following steps:

1. In your Chrome browser: Visit Obs BYGG using this link, and/or Obs by visiting this link
2. The site is in Norwegian, so translate it to English using the google translate feature
3. Select “Our stores” from the menu, then select the store “Alnabru” for Obs BYGG and the store “Arendal” for Obs (This will let you see both the online purchase and the click & collect buttons)
4. At the bottom of your screen you’ll see a toolbar where you can force the treatments



5. In the dropdown you can select between original (24 px radius), which we are calling the treatment, Variation 1 (square/No radius), which is the Control, or Variation 2 (10 px radius), which isn’t used for the throwdown, but will be analyzed.
   If you also set “Force display” to “On”, you’ll be 100% sure you’ll see the selected treatment from the dropdown. You will see the buttons change

⇨ To enter your 1st prediction, visit https://bit.ly/ABThrowdown1

Appendix C – Coupon Code Removal

A description of the motivation for this pattern is in https://experimentguide.com Chapter 2, and in Session 2 of the online course: https://bit.ly/ABClassRKLI.

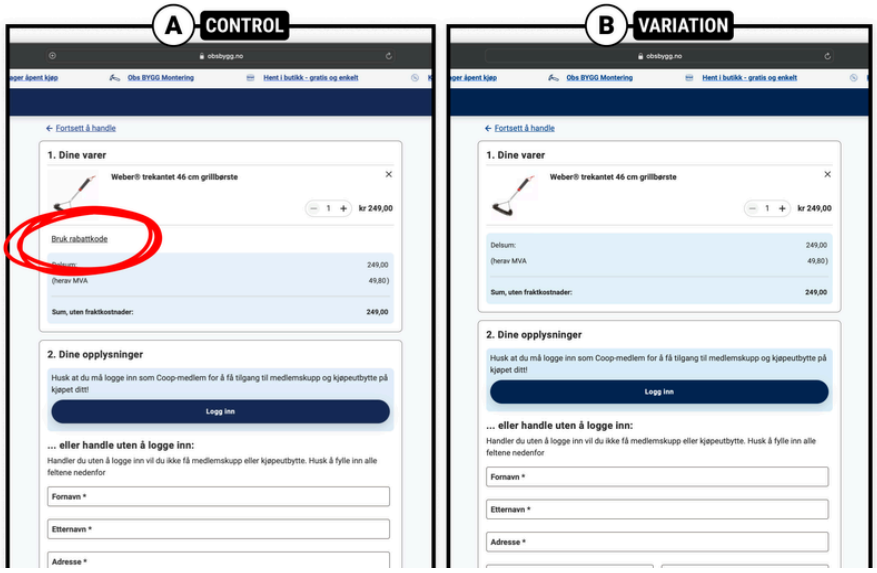
The OEC is conversion rate from initiating checkout to purchase (Boolean per user), again a weighted average of the two sites. The experiment triggers only for users who start checkout and can see the coupon code.

The following table shows prior results, which could help the prediction.

Source	Lift	Comments
Trustworthy Online Controlled Experiments : A Practical Guide to A/B Testing (https://experimentguide.com) Chapter 2, and https://bit.ly/ABClassRKLI	2.8% lift to revenue/user	
GoodUI pattern #1 Only the last one removed the gift card & coupon code; others made the coupon code less visible (e.g., clickable element, not field) (https://goodui.org/patterns/1)	-1.6% to sales 3.3% to sales 2.6% to sales 0.8% to sales 2.4% to sales 24% to revenue	33K visits, p-value 0.07 20K visits, p-value 0.29 15.5K visits. p-value 0.01 3.8K visits, p-value 0.61 4.1K visits, p-value 0.21 No details on number of users or p-value
Evidoo pattern 136 changed coupon to text (lower attraction) for mobile (https://www.evidoo.io/best-practices/136)	+4.0%	63K visitors: There were five A/B tests from electronics, B2C gaming with 40% win ratio, 40% loss ratio

The two variants of interest are as follows:

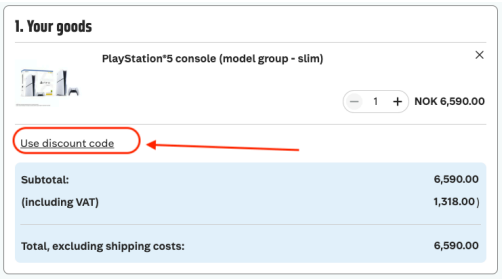
https://bit.ly/ABThrowdown

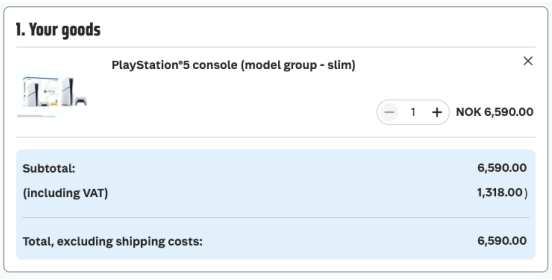


Note that the “Bruk rabattkode,” which means use discount code, is missing in the second. We will also point out that this is not the “in your face” discount code where there is a field that takes up space and attracts the eye, as some of the examples shown in goodUI.

To try the variants live, follow the following steps:

1. Open up incognito mode in your Chrome browser
2. Visit a product page on Obs BYGG here or Obs here.
3. The site is in Norwegian, so translate it to English using the google translate feature
4. Click the “Buy” button (Green for Obs and Orange for Obs BYGG)
5. The item is now in the cart
6. Click on the cart icon in the top right corner
7. Click “To the checkout”
8. You’re now in the checkout. You will see one of these two variants (random at 50% each):





If you get the left (control), you can click on the link and enter a coupon code.
On the right there is no place to enter a coupon code.

9. Repeat the above steps (new incognito window) until you see both variants.

⇨ To enter your 2nd prediction, visit https://bit.ly/ABThrowdown2