

Initial version: 6/2024; last update: 3/30/2026

1. Expert volunteers: [Ronny Kohavi](#), [Jakub Linowski](#), [Lukas Vermeer](#).  
We provide our time, free of charge, to help design, preregister, execute, and analyze the experiment, in exchange for being able to share results that are approved by the companies online and in papers.
2. We are interested in companies that are willing to participate in A/B testing the patterns below. If you are interested, please contact one of us (e.g., Ronny Kohavi) over LinkedIn.
3. Benefits to companies
  - a. Help in executing these A/B tests by the above named experts.  
Feedback and support from experts on design and analysis of experiments.
  - b. Some patterns can have significant ROI for the companies.
  - c. Early access to results. Companies that share results will get access to the other results being analyzed. A paper is likely to take months to publish.
  - d. Improved brand. The participating companies will be acknowledged while the experiments are happening and one person from each company will be a co-author on the paper we will draft and submit. All participating companies will be acknowledged.
4. Requirements from participating companies.
  - a. The expert volunteers will each sign a mutual Non-Disclosure Agreement (NDA) that prohibits sharing any information.
  - b. A separate specific disclosure document will be signed, where a “Draft Result” will be proposed, and the companies can approve it, or propose revisions, or terminate the agreement. The key paragraphs from the disclosure agreement are provided below.
5. We are focused on replicating seven patterns below (some added over time).  
These are strong patterns that are relatively easy to introduce, so companies should be willing to try them and benefit from the analysis. Our goal is to understand the median relative treatment effects and their distribution, and perhaps conditions for when the effects are larger/smaller.

The first two summary papers are

1. <https://bit.ly/powerAnalysisEssentialEJW> - published in Econ Journal watch and focused on the rounded vs. square pattern (abstract at <https://bit.ly/powerAnalysisEssentialEJWAbstract>).
2. <https://bit.ly/trustworthyABPatternsSummary> - summary of several replications that ran until Feb 2026. Will be presented at KDD 2026

The patterns are:

- a. Rounded/square corners – published result that some believe is highly exaggerated.

See paper: <https://bit.ly/powerAnalysisEssential>

Proposed MDE for power calculations on conversion metric (e.g., CTR or conversion to purchase): ~~2%~~ Based on the results of the Coop experiment from January 2025 (<https://bit.ly/ABThrowdown1ResultsPost>), we recommend an MDE lower than 0.56%,

<https://bit.ly/trustworthyABPatterns>

say 0.5% [Updated May 6, 2025]. As was said in Jaws, [we're going to need a bigger boat](#) (a [famous line that wasn't in the script](#))

Historical posts

1. [Initial post on rounded shapes](#) .
2. [SeaWorld result from 7/16/2025](#)
3. [Small Telescopes](#) study

b. Coupon code

[Updated 6/16/2025] Proposed MDE for power calculations on revenue/user: 2% and for conversion: 1% based on data from recent replications.

~~Proposed MDE for power calculations on conversion metric: 2% for lowering prominence (70% of 2.8% below), 4% for removing field (70% of 7.8% would be 5.5%, but is high).~~

The book <https://experimentguide.com> Chapter 2, which was based on a real experiment showed 2.8% and 7.8% for revenue/user, where the former was based on reducing prominence and the latter complete removal.

[GoodUI has five tests](#), all small and borderline: 16K visits, p-value 0.07 negative 1.6% sales; 10K visits, p-value 0.29 positive 3.3% sales; 8K visits, p-value 0.01 positive 2.6% sales; 2K visits, p-value 0.61 positive 0.8% sales; 2K visits, p-value 0.21 positive 2.4% sales; unspecified VWO blog: positive 24% revenue.

<https://www.evidoo.io/best-practices/136/> changed coupon to text (lower attraction) for mobile with the following statistics based on a total of 88K visitors [updated 1/23/2026]:

- i. There were seven A/B tests from Electronics, B2C, Fashion & Shoes
- ii. 42.8% win ratio
- iii. Average treatment effect +3.95%

<https://www.evidoo.io/best-practices/567/> for desktop has the following statistics

- i. There were eight A/B tests from Electronics, B2C, Fashion & Shoes
- ii. 37.5% win ratio
- iii. Average treatment effect +2.2%

[Updated 6/24/2025] Dan's deals shows how Amazon changed their coupon code handling:

<https://www.dansdeals.com/shopping-deals/amazon/amazon-has-changed-how-to-apply-coupons-here-is-e>

<https://bit.ly/trustworthyABPatterns>

[everything-you-need-to-know/](#)

c. Page load time (performance)

Proposed MDE for power calculations: 1% for 250 msec slowdown (70% of  $0.6\% \times 250 / 100$  msec) if done on all pages.

[Updated 3/2025] Lower MDE of 0.2% if done for a page (e.g., just search).

Detailed in Chapter 5 of <https://experimentguide.com>: 100 msec slowdown had 0.6% impact to revenue. Recording of talk:

<https://vwo.com/events/convex-2019/sessions/test-user-experience-bing/>

Summary of results: <https://bit.ly/PerfABPattern>

d. Sticky Call To Action

Proposed MDE: 1% (updated 3/12/2026).

Previously 2.5% (70% of the 3.6% claimed by Evidoo).

See summary at <https://bit.ly/StickyCTAABPattern>

e. Authentic Product Photos

Proposed MDE: 2% (updated from 4% 3/12/2026).

Data is lacking here. 70% of the 5.6% below is 4%. Given the cost to implement something like this, you need a big boost to justify the effort.

Evidoo has <https://www.evidoo.io/best-practices/300> for contextual imagery, and claims

- i. 2 A/B tests from Fashion & shoes
- ii. 100% win ratio, but only 233K users for two experiments on mobile
- iii. Average treatment effect for mobile transactions was 13%

Evidoo has <https://www.evidoo.io/best-practices/231> for AI models and claims

1. One A/B test from fashion & shoes
2. 100% win ratio (for one) with only 233K users
3. Average treatment effect for mobile transactions was 5.6%

<https://goodui.org/patterns/30/> shows three tests, but the largest has just 12K visits and a p-value of 0.09

f. Open link in new tab / open search results in a new tab

See

[https://www.linkedin.com/posts/ronnyk\\_experimentguide-abtest-openinnewtab-activity-7272140955506339841-qGsb](https://www.linkedin.com/posts/ronnyk_experimentguide-abtest-openinnewtab-activity-7272140955506339841-qGsb) for a summary of results

<https://bit.ly/trustworthyABPatterns>

- i. MSN UK home page: 8.9% increase to clicks/user to users triggered by clicking on hotmail (but this is triggered).
- ii. MSN US home page: replication run, similar to above
- iii. MSN Search: +5% increase to clicks/user
- iv. Airbnb search: 3-4% increase to bookings
- v. Sunweb: +7.6%
- vi. GoodUI/job site: +6.4% to successful applications
- vii. GoodUI/VivaReal: +2.2% to form submissions

Proposed MDE: 2%

This is a trivial change and one of the most successful patterns we have ever seen.

It is a trivial change: one only needs to add: target="\_blank" as explained in

[https://www.w3schools.com/tags/att\\_a\\_target.asp](https://www.w3schools.com/tags/att_a_target.asp).

Note that this pattern is likely more useful for desktop/web than mobile. For example, at Airbnb, it's only for desktop/web.

- g. Underline links.

See <https://bit.ly/underlineLinksPost> and <https://bit.ly/underlineLinks>.

There are currently no A/B tests about this in GoodUI or in Evidoo.

Proposed MDE: 1% for a key metric like add-to-cart, purchase conversion rate, or revenue per user.

Note that for click-through rate, it is likely easy to increase clicks on some areas by adding underlines to those, but it hurts consistency, and may shift clicks from other places.

There are other variants that could be tried, as documented by Jakub at

[https://www.linkedin.com/posts/jlinowski\\_experimentation-exploration-ui-activity-7414697431147163648-amWD](https://www.linkedin.com/posts/jlinowski_experimentation-exploration-ui-activity-7414697431147163648-amWD)

## 6. Pre-registrations (added 1/18/2026)

- a. OSF pre-registrations are at <https://osf.io/gen6a/registrations>
- b. Subjects to use in pre-registrations:
  - i. Business->E-Commerce
  - ii. Psychology -> Experimental Analysis or Behavior
  - iii. Physical Sciences and Mathematics -> Statistics and Probability -> Applied Statistics
  - iv. Physical Sciences and Mathematics -> Statistics and Probability -> Clinical Trials
  - v. Physical Sciences and Mathematics -> Statistics and Probability -> Design of Experiments
  - vi. Psychology -> Applied Behavior Analysis
- c. Tags to use:
  - i. A/B test
  - ii. Trustworthy A/B Patterns

7. Meta-analysis (added 9/26/2025)

Given concerns by Data Colada about meta-analysis shared in <https://datacolada.org/73> and [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3271372](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3271372), we pre-declare our process for meta-analysis as follows

- a. For each pattern, we will include all trustworthy analyses in the meta-analysis.
- b. An analysis is NOT trustworthy if it has red flags, such as an SRM with p-value < 0.05 (normally < 0.001 in some tools, but we prefer to be conservative), if there is significant user contamination, where over 5% of users were exposed to multiple treatments, or where the data logged is deemed unreliable. We realize that we are allowing some human degrees of freedom to determine “red flags.”  
When in doubt about the data quality, we exclude the replication run.
- c. For meta-analysis, we will use the non-CUPED version (<https://bit.ly/expCUPED>) of experiment results due to technical difficulties in doing meta-analysis (the meta-analysis tools require the success matrices and tools don’t support sufficient CUPED reporting)
- d. Experiments that did not follow the pre-registration protocol, such as minimum duration for required power, will be excluded

8. FAQ (from meetings)

- a. Does a participating company need to run all patterns, or can they choose?  
They can choose. Given the overhead of joining the project, we would love to see companies running a few patterns, not just a single one.
- b. What is the timeline?  
We recognize that planning/designing/running/analyzing/certifying takes time.  
We don’t have a pre-set timeline, but we assume this will take several months.  
This is our V1, and when we have sufficient replications for an interesting result to share, we might drop some planned runs, and perhaps do round 2.
- c. Does disclosing sample size not reveal confidential information not otherwise available?  
Because we are not specifying the experiment runtime (e.g., the experiment could have run for one week or eight weeks), the sample size provides very limited information.  
In addition, we do not specify that the experiment was run on 100%, or was triggered to a smaller sample. For example, an A/B pattern could be tested as an A/B/C experiment, with A/B our goal pattern, but C is a related pattern that a company wants to test, so we’re effectively running on 66% of traffic. All these things make estimates of traffic from these results imprecise.
- d. Are there examples where companies published experiment results?  
Microsoft (where Ronny Kohavi worked) published multiple experiments:
  - [https://bit.ly/HBR\\_AB](https://bit.ly/HBR_AB)
  - <https://bit.ly/expRulesOfThumb>

Booking (where Lukas worked) published multiple experiments:

- <https://www.ueo-workshop.com/wp-content/uploads/2014/04/noulasKDD2014.pdf>

<https://bit.ly/trustworthyABPatterns>

- <https://dl.acm.org/doi/abs/10.1145/2766462.2776777>
- <https://dl.acm.org/doi/10.1145/3292500.3330744>

GoodUI.org (headed by Jakub) has multiple experiments

- <https://goodui.org>

Vivli has 56 companies that pay to share their information about all clinical trials. See <https://bit.ly/vivliRCTSharing>

An earlier version of this doc is available [here](#).

## Appendix: Exaggerated Claims

How many believe this ad?



SPONSORED CONTENT / By The Skincare Magazine

**56-Year-Old Replaced \$21,500 Facelift With This \$38 Drugstore Item** [↗](#)

**Claim:** a \$39 “drugstore serum” is a “**facelift in a bottle,**” that “**erases 99.7% of wrinkles... within 15 hours,**” and delivers “**identical results to \$12,000 facelifts.**”

This liquid solution



It leads to <https://dailyskinmagazine.com/flawless-skin-over-50pbbs-h4-b-us-otb> with a claim that it “erases 99.7% of wrinkles... in 15 HOURS.” That’s far beyond what topical skincare can do.

<https://bit.ly/trustworthyABPatterns>

The **ONLY** drugstore serum being called "**FACELIFT IN A BOTTLE**" that **FILLS**, **FIRMS**, and **ELIMINATES** even the **DEEPEST** wrinkles, fine lines, and sagging skin in **just 15 HOURS**.



Gemini wrote: **Implausible Speed:** It claims a "powerful tightening effect is visible... within **15 HOURS**" and that deep wrinkles "**VIRTUALLY DISAPPEARED**" overnight, which is far faster than the typical results of clinically proven skincare ingredients.

Only \$50 at <https://try.olavita.co/checkout-special-offer-pr-us-otb-st2>

☐ **1 Bottle** - Save 50%



Reg ~~\$100~~  
**\$50 each**  
Savings \$50  
FREE SHIPPING

☒ **3 Bottles** - Save 62%



★ **Most Popular**  
Reg ~~\$300~~  
**\$38.95 each**  
Savings \$185  
FREE SHIPPING

<https://dailyskinmagazine.com/clinicalstudies> - they mis-spell the title :-)

CLINICAL STUDIES

## Appendix: Disclosure Agreement

Here are the two key paragraphs of the disclosure agreement.

**WHEREAS**, Kohavi and his colleagues, including Lukas Vermeer and Jakub Linowski, have each entered into a confidentiality, non-disclosure, or similar agreement with Company (the “**NDAs**”) in connection with the Project and the development, implementation, evaluation, and other consulting related to certain A/B tests to be conducted by the Company (the “**Company A/B Tests**”); and

**WHEREAS**, Kohavi and Company desire to publish certain summary statistics, insights, and aggregated data generated or learned from the Company A/B Tests (the “**Published Results**”), subject to the terms and conditions of this Agreement.

**NOW, THEREFORE**, in consideration of the mutual promises and covenants herein contained, the Parties agree as follows:

1. Publication Review. From time-to-time during the Term, Kohavi will provide to Company a proposed draft of the Published Results that Kohavi proposes be made publicly available (“**Draft Results**”). Within fourteen (14) days following its receipt of the Draft Results, Company will, acting in good faith, either: (i) approve publication of the Draft Results or (ii) propose revisions or other comments to the Draft Results. If Company approves the Draft Results, Kohavi and his designees may publish and otherwise publicly disclose the Draft Results and such Draft Results will be considered Published Results hereunder and available to the public for purposes of, and not otherwise subject to, the NDAs. If Company proposes revisions or comments to the Draft Results, Kohavi may revise the Draft Results and resubmit them for Company’s review in accordance with the process set forth in this Section 1.
2. Scope of Published Results. The Parties currently anticipate that the Published Results would include the following information, except to the extent such information is competitively sensitive or otherwise subject to a *bona fide* confidentiality concern:
  - a. Company’s identity as a participant in the Project;
  - b. Company’s identity as the source of the applicable Company A/B Test;
  - c. Description of the applicable Company A/B Test with a level of detail sufficient to enable replication from such description by an outside observer (e.g., screenshots of control and treatments);
  - d. For each relevant metric tested by the applicable Company A/B Test, the relative treatment effect (e.g., lift), confidence interval, p-value, sample size, and the statistical power with such metric’s associated relative minimum detectable effect (MDE);
  - e. Summary statistics related to trust in the applicable Company A/B Test, such as sample-ratio and caveats related to trust (e.g., logging issues); and
  - f. Subject to the Company’s reasonable approval, such other notable or differentiated segments and the impact of such segments on relevant metrics tested by the applicable Company A/B Test.