~~https://bit.ly/trustworthyABPatterns~~

V2 - **OLD proposal **

Last update: 11/28/2024 with examples of publications in FAQ

1. Name: A/B Patterns Reproducibility project
2. Expert volunteers: Ronny Kohavi, Jakub Linowski, and Lukas Vermeer, and the Center for Open Science.
   We provide our time and effort to help design, execute, and analyze the experiment, in exchange for being able to share results that are approved by the companies online and in papers.
3. Companies and/or vendors will be asked to participate.
4. We will focus on five patterns
   a. Rounded/square corners – published result that some believe is highly exaggerated. See https://www.linkedin.com/posts/ronnyk_do-elements-with-rounded-shapes-enhance-click-through-activity-7183554027773734914--ugr

      Proposed MDE for power calculations on conversion metric: 2%.
      This is a tough one to estimate, but several practitioners believe this is much closer to zero than to the 17% plus in the paper.

   b. Coupon code
      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%, 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 0.2.4% 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 63K visitors:

      i. 5 A/B tests from electronics, B2C gaming
      ii. 40% win ratio, 40% loss ratio
      iii. Average treatment effect +4.0%

   c. Page load time (performance)
      Proposed MDE for power calculations: 1% for 250 msec slowdown (70% of 0.6%*250/100 msec).

      Detailed in Chapter 5 of https://experimentguide.com: 100 msec = 0.6% revenue (evaluated at 250 msec).
      Recording of talk:
      https://vwo.com/events/convex-2019/sessions/test-user-experience-bing/
      Requires server-side testing capability.

d. Sticky Call To Action

Proposed MDE: 2.5% (70% of the 3.6% claimed by Evidoo).
https://www.evidoo.io/best-practices/114/ . Evidoo claims:

i. 14 A/B tests (from multiple domains, including Fashion, Furniture, Jewelry, Outdoor, Consumer Goods) with 57% win ratio, 7% loss (rest are flat).
ii. Average treatment effect (purchase/transaction) is 3.6%, based on 3M users in total (over all experiments).

https://goodui.org/patterns/41/ shows 13 tests, but most underpowered.

https://goodui.org/patterns/41/tests/217/ had 80K visits with increase to engagement, tiny p-value below 0.001, but this was not conversions.

e. Authentic Product Photos

Proposed MDE: 4% (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%

Evido 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

The idea is that 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 perhaps conditions on conditions when the effects are larger/smaller.

5. Benefits to companies
   a. Help in executing these A/B tests.
   b. Feedback and support from experts on design and analysis of experiments.
      Ron Kohavi, Lukas Vermeer, Jakub Linowski, people from the Center for Open Science, and others will provide feedback.
      To see the value of this, have a look at the post on Vivli: https://bit.ly/vivliRCTSharing

   c. Early access to results for months. 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.
6. Requirements from participating companies. There are two levels:
   a. Proposal level: companies will document which patterns they are interested in running and will sign an NDA, allowing information sharing with the "experts" for the design of the experiment, but excluding the experts from sharing information outside the limited group. We will also sign a two-page *disclosure* document detailing our expectation that the companies will share data about the result. The companies will review a draft of the disclosure and can approve or modify it before anything is publishable.
   b. Post-experiment level: companies that have run the experiment can choose to share the results at two levels
      i. [Required] Summary-only: number of users, treatment effects for key metrics, p-values for key-metrics, and confidence intervals for key metrics, and optionally these for interesting segments.
      ii. [Optional] Raw-data: raw data will be shared only with the experts for deeper analysis/feedback.

   Summaries will be jointly written and approved by the companies before being shared. We envision two levels

   a. Summary to be shared among participating companies only.
      Companies will be able to specifically include/exclude companies for competitive reasons (e.g., Booking may decide that their detailed results should not be shared with Airbnb).
   b. Summary to be shared publicly, that is, draft for the paper that will be written.

FAQ (from meetings)

1. 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.
2. 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.
3. 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 8 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.

4. Are there examples where companies published experiment results?
   Microsoft (where Ronny Kohavi worked) published multiple experiments:
   - 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://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

V1: Trustworthy A/B Testing Patterns – Initial Ideas

Ronny Kohavi
6/5/2024

TL;DR: Can we create a corpus of reliable patterns for websites / mobile applications that replicate with high probability? The Center for Open Science has shown that many results in Psychology do not replicate (https://doi.org/10.1126/science.aac4716) and it's clear that the same is true in the online world.

Background

Online controlled experiments, or A/B tests, are relatively easy to conduct: the necessary ingredients that make offline testing (e.g., in medicine) hard, are all accessible online (see https://bit.ly/expBookChapter1):

1. Randomization can be done reliably using hash functions, so there are no issues with blinding of assignments (e.g., biases can be introduced during human assignment to variants).
2. Data is collected electronically (although there are instrumentation bugs sometimes).
3. Data is plentiful – many sites have access to hundreds of thousands or millions of users.
4. Motivation is high: impact to the business is highly material. The opening example in http://experimentguide.com shows a 12% increase to revenue for Bing, which was worth over $100M at the time, and now worth five times that.

5. Most ideas fail to improve the metrics they were designed to improve. Only about 10-15% of online experiments are statistically significantly positive (See https://bit.ly/ABTestingIntuitionBusters Table 2).
6. Scale: large organizations run tens of thousands of A/B tests per year. See https://bit.ly/OCESummit1

The biggest problem with A/B tests is that effect sizes are small necessitating large sample sizes. In the Fat Tails paper (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3171224), Table 1 shows that on a key Bing metric, success rate (Boolean), the max treatment effect was 0.28% from over 1,000 experiments.

Existing Resources

There have been several attempts to share interesting patterns:

1. We wrote a paper called Rules of Thumb (https://bit.ly/expRulesOfThumb) with patterns, such as: performance has a large impact on key metrics; reducing abandonment is hard because you usually just shift clicks.
2. Optimizely published a summary of 127K experiments: https://www.linkedin.com/posts/ronnyk_lessons-learned-from-running-127000-experiments-activity-7143376795940106240-faUh
3. Jakub Linowski's website: https://goodui.org is intended to share patterns and A/B tests. GoodUI publishes all submissions, so there is no quality control, and there is a mix of good and bad experiments in the hope that this will be self correcting.
4. Deborah O'Malley publishes a bi-weekly test at https://guessthetest.com/ . Given the pressure to send something every two weeks, the quality varies. There is a nice trustworthy section to result, attempting to quantify the trust level. .

There are a lot of bad results out there, which I've given the Twyman's Law awards to: .

1. Some are just statistically naïve, such as https://bit.ly/TLAward1
   This person has a huge audience of 463K followers, and claimed 39.7% improvement based on 6 conversions out of 154 visitors rising to 8 conversions out of 147 visitors.
2. Another example discussed in https://bit.ly/TLAward2 showed 24% improvement based on 37 sessions. The original post was removed, so the criticism achieved its goal.
3. Some actually got published in journals with dubious results:https://bit.ly/TLAward3 .
   Here is my analysis:
   https://www.linkedin.com/posts/ronnyk_do-elements-with-rounded-shapes-enhance-click-through-activity-7183554027773734914--ugr
4. Some of the examples in https://guessthetest.com are under-powered. I criticized an extreme example at https://bit.ly/ABTestingIntuitionBusters

~~https://bit.ly/trustworthyABPatterns~~

Proposal – initial ideas

1. Try to replicate what COS did with reproducibility of psychological science https://www.science.org/doi/10.1126/science.aac4716
2. A key difference: the pattern will be useful for many web sites.
   I'm thinking of a reliable version of https://goodui.org, perhaps collaborating with Jakub.
3. Work with A/B testing vendors to select patterns and replicate them.
4. Key question: what's in it for them?
   a. Co-authorship on "the" paper: individual or company.
   b. Free help from experts like me and others in COS
   c. Contribution to the science of trustworthy A/B tests
   d. Increased odds of success (from surfacing positive probability patterns for replication)
5. Could also approach companies directly, although the big ones will see more risk than value.
6. Key concerns: disclosing sensitive data.
   a. Come up with a legal agreement that would satisfy companies.
      For example, no absolute metrics, only relative improvements.
   b. Ability to veto or modify text
   c. NDA with experts like me and COS
7. Do a trial run with the rounded/square buttons?
   https://www.linkedin.com/posts/ronnyk_do-elements-with-rounded-shapes-enhance-click-through-activity-7183554027773734914--ugr
8. COS involvement. In initial discussions with Brian Nosek, he thought we might do it as part of their SMART project (https://www.cos.io/blog/join-the-smart-project-advancing-automated-research-evaluation). SMART is a Robert Wood Johnson Foundation funded project to build on the DARPA SCORE project to develop algorithms to assess the credibility of research findings. One component of SMART is subjecting claims from a subset of papers to replication. Algorithm teams rate the credibility of the claims from the entire set of papers.
9. Pre-registration. All experiments will have to be pre-registered with the requirement to publish the result, which could be "invalid experiment." This is different from sites, such as https://aspredicted.org/messages/private_forever.php, which allow keeping pre-registrations private forever, as it defeats the purpose (I could register 20 and publish the one stat-sig).
   For example, https://bit.ly/TLAward3 (rounded corners paper) pre-registered their experiments with https://aspredicted.org, but we don't know how many pre-registered experiments were not published.
10. [Jakub Linowski proposed this] Some factors which may increase success rates (accuracy?) of these replications:
    a. Pre-registration (to minimize publication bias) [Agree, see above. RonnyK]
    b. Low or estimate based MDEs (some patterns might need lower/higher MDEs) [Determining the power is part of the pre-experiment review]
    c. Covariates (where does it work? not work?), that is, are there interesting segments where the effect seems larger (HTE).
    d. Change intensities (how pronounced does the change need to be to be effective? Does it have limits?)

    e. Confounders & interactions (does it work better/worse with other ingredients/variables?)
    f. Consistency of metrics (some patterns may come with trade-off situations for different metrics)
11. [Jakub Linowski proposed this] Track both predictions or estimates and actual effects (to track which types of approaches improve reproducibility; which don't; and improve over time) Kellen Mrkva suggests polling predictions from CRO experts. She points to DellaVigna and Milkman, which do that in their research on nudges and government RCTs using predictions through https://socialscienceprediction.org/.