Lessons from Running Thousands of A/B Tests

Ronny Kohavi
Distinguished Engineer, GM, Analysis and Experimentation, Microsoft

A/B Test in One Slide

- Concept is trivial
  - Randomly split traffic between two (or more) versions
    - A (Control)
    - B (Treatment)
  - Collect metrics of interest
  - Analyze
- Sample of real users, not WEIRD (Western, Educated, Industrialized, Rich, and Democratic) like many academic research samples
- A/B test is the simplest controlled experiment
- Must run statistical tests to confirm differences are not due to chance
- Best scientific way to prove causality, i.e., the changes in metrics are caused by changes introduced in the treatment(s)
Simple Experiments, Large Scale

- We run ~300 concurrent experiments at Bing on a given day
- Each experiment typically involves 100K to millions of users
Example: Bing Ads with Site Links

Should Bing add “site links” to ads, which allow advertisers to offer several destinations on ads?

OEC: Revenue, ads constraint to same vertical pixels on avg

A

Pro: richer ads, users better informed where they land
Cons: Constraint means on average 4 “A” ads vs. 3 “B” ads
Variant B is 5msc slower (compute + higher page weight)

B

• Raise your Left hand if you think A Wins
• Raise your Right hand if you think B Wins
• Don’t raise your hand if you think they’re about the same
Bing Ads Example

- If you raised your left hand, you were wrong
- If you did not raise a hand, you were wrong
- Site links generate incremental revenue on the order of tens of millions of dollars annually for Bing

The above change was actually costly to implement. But we made two small changes to Bing, which took days to develop, each increased annual revenues by about $100M (One was delayed by 6 months because it was not prioritized high, a prioritization mistake that cost $50M)

Several examples in our [KDD 2014 paper](#)
1. Assessing the value of novel ideas is hard
2. The OEC (Overall Evaluation Criterion) is critical
3. There are never enough users
4. Getting numbers is easy; getting numbers you can trust is hard!
Assessing the Value of Novel Ideas is Hard

- We are often fooled by correlation.
- Doctors did bloodletting for 2,000 years.
  - Bloodletting has a calming effect on patients.
  - Through the 1830s the French imported about forty million leeches a year for medical purposes.
  - President George Washington had a sore throat. Three doctors extracted 35% of his total blood in one night, causing anemia and hypotension.
  - He died that night.
- Ioannidis evaluated the reliability of forty-nine influential studies (each cited more than 1,000 times).
  - 90 percent of large randomized experiments produced results that stood up to replication, as compared to only
  - 20 percent of nonrandomized studies.
Assessing the Value of Novel Ideas is Hard (2)

- Features are built because teams believe they are useful. But most experiments show that features fail to move the metrics they were designed to improve.
- Based on experiments at Microsoft (paper), 2/3 of ideas evaluated using controlled experiments were flat or negative.
- At Bing, which is well optimized, failure rate is about 80%-90%. We joke that our job is to tell clients that their new baby is ugly.
- In the book *Uncontrolled*, Jim Manzi writes
  - [At Google] only about 10 percent of these leading to business changes.
- In Experimentation and Testing Primer by Avinash Kaushik, he wrote 80% of the time you/we are wrong about what a customer wants.
Learnings from First Lesson

Avoid the temptation to try and build optimal features through extensive planning without early testing of ideas

Experiment often

- *To have a great idea, have a lot of them* -- Thomas Edison
- *If you have to kiss a lot of frogs to find a prince, find more frogs and kiss them faster and faster*  
  -- Mike Moran, *Do it Wrong Quickly*

Try radical ideas. You may be surprised

- Doubly true if it’s cheap to implement
- *If you're not prepared to be wrong, you'll never come up with anything original* – Sir Ken Robinson, TED 2006 (#1 TED talk)
Lesson 2: the OEC

If you remember one thing from this talk, remember this point:

OEC = Overall Evaluation Criterion

- Agree early on what you are optimizing. It’s HARD!
- Getting agreement on the OEC in the org is a huge step forward.
- Suggestion: optimize for customer lifetime value, not immediate short-term revenue.
- Criterion could be weighted sum of factors, such as:
  - Visits (e.g. Sessions/user)
  - Revenue/user (under some constraints)
  - Success per visit (however success is defined)
  - Time to success (faster is better) or time on site
- Report many other metrics for diagnostics, i.e., to understand the why the OEC changed and raise new hypotheses.
Lesson 3: There are Never Enough Users

Assume a metric of interest, say revenue/user
  - Denote the variance of the metric by $\sigma^2$
  - Denote the sensitivity, i.e., the amount of change we want to detect by $\Delta$

From statistical power calculations, the number of users ($n$) required in experiment is proportional to $\sigma^2 / \Delta^2$

The problem
  - Many key metrics have high-variance (e.g., Sessions/User, Revenue/user)
  - As the site is optimized more, and as the product grows, we are interested in detecting smaller changes (smaller $\Delta$)

Example: A commerce site runs experiments to detect 2% change to revenue and needs 100K users per variant.
For Bing US to detect 0.1% ($2M/year), we need $20^2 \times 100K = 40M \times 2$ variants = 80M users (Bing US has about 100M users/month)
We must run large experiments

- Bing runs 10-20% per variant, and sometimes 45/45% (we keep a 10% global holdout). Most sites should be running 50%/50% experiments.
- Users are now in multiple concurrent experiments (see Large Scale paper).

Use variance reduction techniques

- **Triggering**: analyze only users who were actually exposed to change.
- Use lower-variance metrics (e.g., trim revenue, or look at Boolean metrics like conversion rate vs. revenue; see paper Section 3.2.1).
- Use pre-experiment period: before the experiment started, there was no difference between the control and treatment. We can use the deltas in the pre-experiment period to reduce the variance. Nice trick called CUPED.
- Reduce impact of chance by rejecting randomizations that fail the pre-experiment A/A test (see paper Section 3.5).
Lesson 4: Getting numbers is easy; getting numbers you can trust is hard!

- There is a saying that
  
  *The difference between theory and practice is larger in practice than the difference between theory and practice in theory*

- Enormous amount of time needs to be spent on data quality
  
  - Running a series of A/A tests typically shows failure rates much above the expected 5% (e.g., 30% failures on new sites).
  - The most common reason is carryover effects (see Section 3.5 of paper)
  - Sample ratio mismatch (getting 49/51% on a 50/50% design) is our most common signal that something is terribly wrong
  - Click instrumentation is either reliable or fast (but not both; see paper)
  - Bots can cause significant skews. At Bing over 50% of traffic is bot generated

**Twyman’s law**: *Any figure that looks interesting or different is usually wrong*

- See Pitfalls, Puzzling Outcomes, Practical Lessons
Assessing the value of novel ideas is hard
- Prepare to be humbled. When ideas are objectively evaluated with controlled experiments, the failure rate (flat or worse) is 60-90%
- Culturally, it is hard to change from “Do x” to “Evaluate x,y,z”

The OEC (Overall Evaluation Criterion) is critical
- Make sure you agree what the org is optimizing for. It is HARD!

There are never enough users
- Detecting small differences requires large experiments (e.g., 50/50%)
- Utilize variance-reduction techniques

Getting numbers is easy; getting numbers you can trust is hard!
- Twyman’s law: Any figure that looks interesting or different is usually wrong