A/B Testing: The Most Powerful Way to Turn Clicks into Customers
By Dan Siroker, Pete Koomen and Cara Harshman

CHAPTER

3

Seek the Global Maximum

Refinement and Exploration

Premature optimization is the root of all evil.

Donald Knuth

Imagine you're climbing a mountain: if your goal is to get to the top of the *tallest* mountain, and you don't have a map of the range, it's probably not a good idea just to start walking up the nearest slope. You'll climb, and climb, and then ultimately reach some peak—and then what? Where do you move next if this peak doesn't turn out to be the highest one?

In optimization, the term for the nearby, uphill peak is the *local maximum*, whereas the distant, largest peak is the *global maximum*.

One of the things that we like to tell companies that we work with is to be willing to *think big*. Being too complacent about the status quo can lead to focusing too much on fine-tuning. As Figure 3.1 highlights, the "Refinement" path might lead you to miss out on the best solution that could have been discovered with the "Exploration" approach. While refinement can lead to a solution better than what you have today, we recommend exploring multiple alternatives that might not resemble the current site first. We encourage the kind of humility and bravery required to say, "You know, the website we have today is far from perfect. Let's try some *dramatically* new layouts, new designs, and redesigns, figure out which of those work well, and *then* refine from there."

However, it's not as simple as saying that one should always explore first and always refine second. The truth is that exploration and refinement are complementary techniques, and most effective when used in tandem. Often the process of using

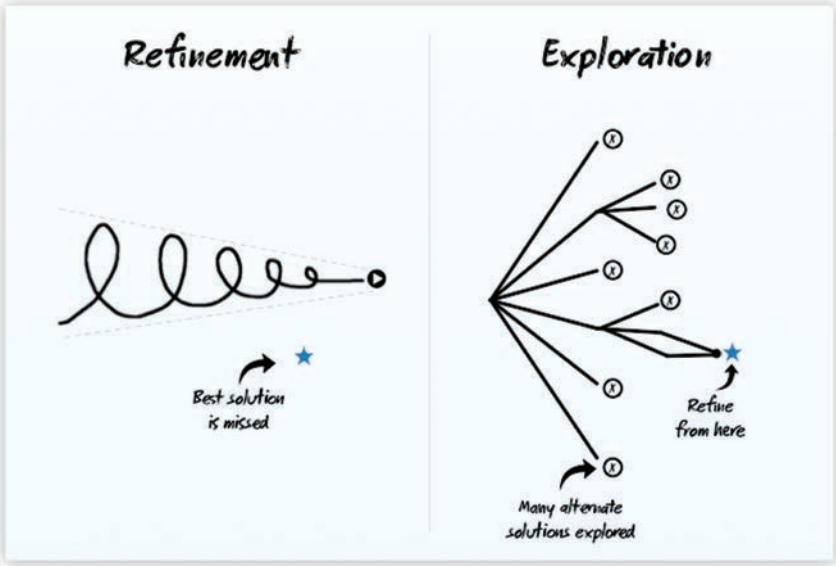

FIGURE 3.1 Refinement and exploration.
Source: Intercom.

hypothesis testing for refinement produces key insights that can deeply inform the redesign. In other words, sometimes you need to get above the tree line to see where the bigger peak lies.

As the following case studies reveal, there are huge wins to be had from thinking big and being open to questioning the status quo. There are also important revelations lurking in smaller tests that can point the way to a major redesign. And sometimes testing is the only way to find true north amidst the chaos and confusion of major changes.

Break from the Status Quo: ABC Family

Disney ran an experiment using Optimizely on the ABC Family homepage.

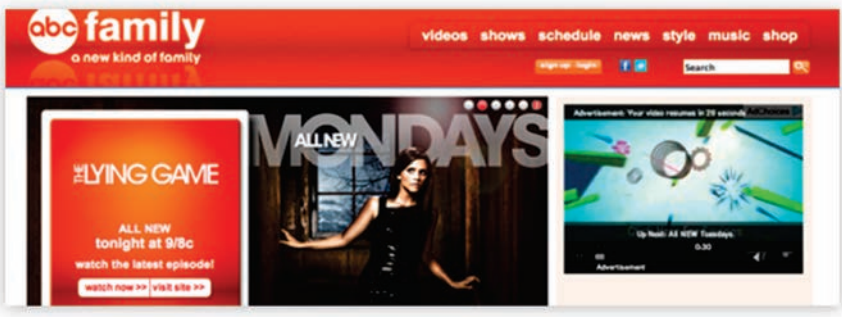

FIGURE 3.2 ABC Family A/B test: original.

The page (shown in Figure 3.2) displayed a large promotion for a television show you might be interested in. After looking through their search logs, however, the Disney digital team discovered that a lot of people were searching for the exact titles of shows and specific episodes. Instead of taking the incremental approach (e.g., by tweaking the promo image, or rotating the featured show), the team decided to reevaluate their entire approach. They created an alternative view, one that was less visual and more hierarchical, in which users can drill down through menus to specific shows (Figure 3.3).

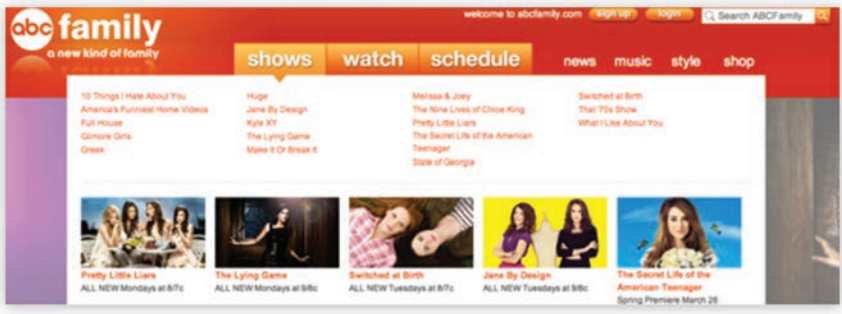

FIGURE 3.3 ABC Family A/B test: variation.

Disney had defined as their quantifiable success metric the percentage of visitors who clicked on any part of the experiment page. Their goal was to lift this engagement by 10 to 20 percent. In fact, by being open to this big, fundamental change, they were able to effect an engagement increase of more than *600 percent*.

Learn Your Way to the New Site: Chrome Industries

Kyle Duford at cycling bag and apparel manufacturer Chrome Industries explains that the Chrome team is presently discussing a major site redesign. "We're purposely using all of these tests to formulate how we approach the new website."

The Chrome team discovered something surprising when they were A/B testing the order of the three promotional content blocks on their homepage: the content they put in the center block seemed *always* to outperform the content they put in the left block (Figure 3.4).

The team's assumption was that because people read from left to right, they would explore in this manner. "This is gold," says Duford. Now they know to put their most important promo block in the center, but the bigger lesson is that users seem to go straight for the central imagery, rather than scanning left to right. This is a valuable insight that may end up altering the entire new layout for the site redesign. "The look and feel will be completely different, but the ideas of the blocks of content that go into it are all being discovered through this process," Duford says. "So while it's important right now to understand how people shop, it's more important because it's going to inform our decisions going forward."

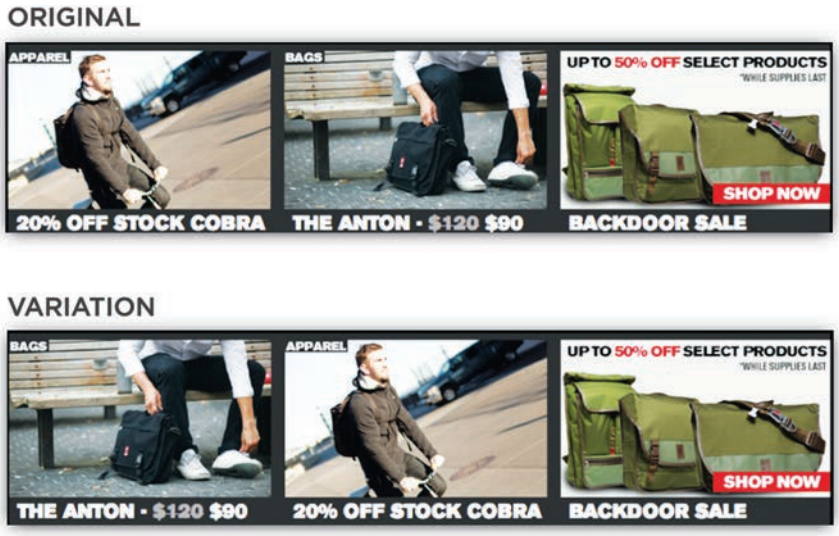

FIGURE 3.4 Chrome Industries promo block test—middle block most clicked, regardless of content.

Rethink the Business Model: Lumosity

Lumosity is a company that offers web-based games designed to improve users' minds. Their business model is simple: users pay a monthly subscription fee to access games designed by neuroscientists to promote cognitive function. Users derive the most benefit from training regularly, and boosting user engagement was an important goal. What wasn't intuitive, however, was what the Lumosity development team did to increase this metric.

Lumosity's scientists recommended that users train for 15 to 20 minutes a day, 4 to 5 times per week—not unlike physical exercise—although the site didn't actually constrain users to a specific time limit. The data showed that people would stay logged in for many hours, but that over time, the frequency of logins declined, suggesting users were burning out.

The team hypothesized that limiting the amount of training a user could do in one day would improve engagement over time.

Giving users one training session a day and congratulating them on being done for the day might achieve their goal. Such a radical change initially made many people at the company nervous, including Product Manager Eric Dorf, who feared that restricting the amount of a time a user could use the service they were paying for would frustrate the user base. "I felt like, if I'm paying for this as a subscription and I'm not allowed to train as much as I want, why would I pay for it?" he says. "I remember thinking, 'Gosh, I hope we don't piss everybody off.'"

Trying out the new model as part of an A/B test mitigated that risk. The team ran an A/B test that set the original, unlimited training against the limited-training variation (Figures 3.5 and 3.6). The results shocked Eric and his team. Users actually trained more over time in the new model. "The graph was so clear," Eric says. "People were training more as a result of being limited."

After making this discovery, the Lumosity team changed the way they position, build, and sell their program. The message of daily training is the cornerstone of their communications to users. After this initial exploration, the team then subsequently used A/B testing to refine the approach, finding the messages and marketing that best support and reinforce the idea of daily training.

Today, when a user completes a session, the message is, "You're done. You can leave the site now," Dorf explains. "It's not like a lot of other gaming products that want you to spend all your time playing. The scientists are happy because more users are more engaged with training than before."

Test *Through* the Redesign, Not After: Digg and Netflix

When it comes to making better data-driven decisions, the sooner the better. Often the temptation is (and we've heard this before)

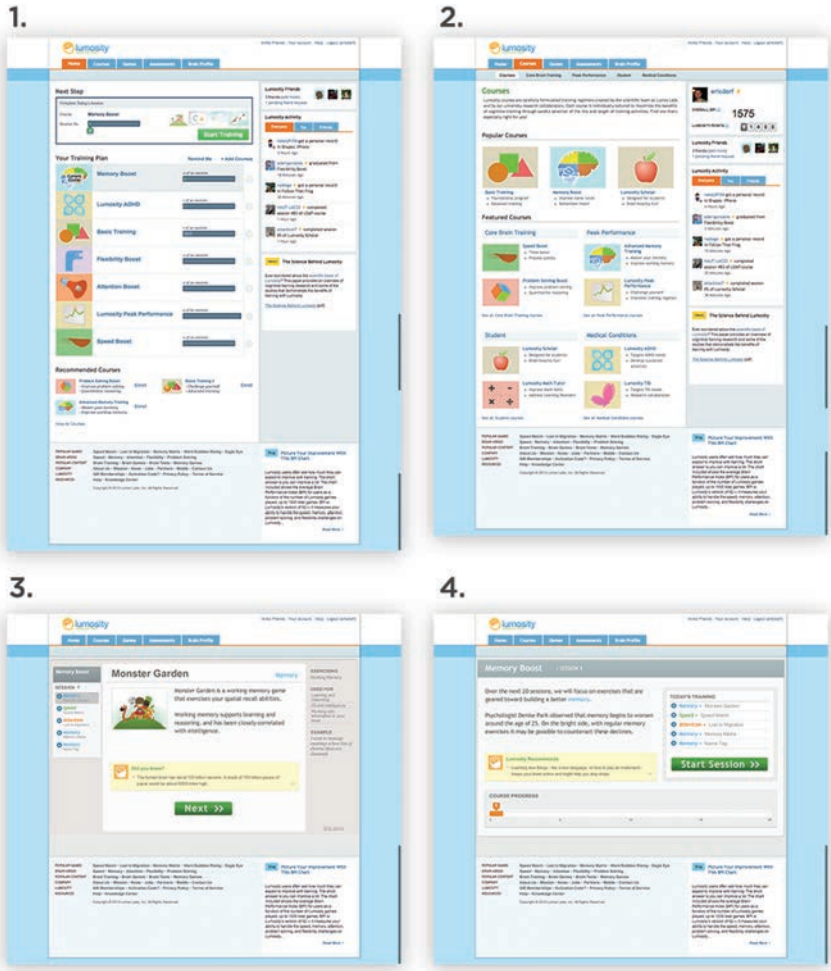

FIGURE 3.5 Original Lumosity user experience—unlimited daily training.

"Oh, we're doing a redesign; we'll do the A/B testing afterwards." The fact is you actually want to *A/B test the redesign*.

Around 2010, we were introduced to the folks at Digg by their new VP of Product Keval Desai to talk about using Optimizely. Their response was, "We are busy working on a complete overhaul of our site. After we do that, then we'll do A/B testing."

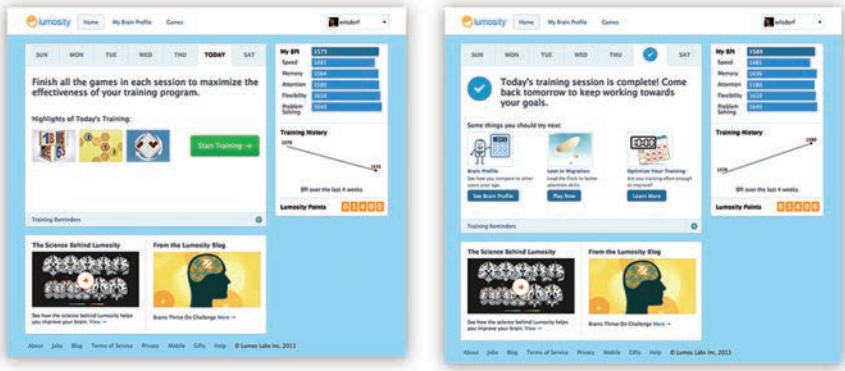

FIGURE 3.6 New Lumosity user experience—limited daily training.

As Desai explains, the "Digg v4" redesign was a perfect storm of problems. The company rolled out a new backend and a new frontend at the same time, conflating two different sets of challenges. "It was a big bang launch," he says. The backend couldn't initially handle the site traffic and buckled on launch day. What's more, despite faring well in usability tests, focus groups, surveys, and a private beta, the new frontend met with vociferous criticism when it was released to the public, and became a magnet for negative media attention. "When you change something, people are going to have a reaction," Desai says. "Most of the changes, I would say, were done for the right reasons, and I think that eventually the community settled down despite the initial uproar." But, he says, "a big-bang launch in today's era of continuous development is just a bad idea." "To me, that's the power of A/B testing: that you can make this big bet but reduce the risk out of it as much as possible by incrementally testing each new feature," Desai explains. People are naturally resistant to change, so almost any major site redesign is guaranteed to get user pushback. The difference is that A/B testing the new design should reveal whether it's *actually* hurting or helping the core

success metrics of the site. "You can't [always] prevent the user backlash. But you can know you did the right thing."

Netflix offers a similar story of a rocky redesign, but with a crucial difference: they were A/B testing the new layout, and had the numbers to stand tall against user flak. In June 2011, Netflix announced a new "look and feel" to the Watch Instantly web interface. "Starting today," wrote Director of Product Management Michael Spiegelman on the company's blog, "most members who watch instantly will see a new interface that provides more focus on the TV shows and movies streaming from Netflix." At the time of writing, the most liked comment under the short post reads, "New Netflix interface is complete crap," followed by a litany of similarly critical comments. The interface Netflix released to its 24 million members on that day is the same design you see today on netflix.com: personalized scrollable rows of titles that Netflix has calculated you will like best. So, in the face of some bad press on the blogosphere, why did Netflix decide to keep the new design? The answer is clear to Netflix Manager of Experimentation Bryan Gumm, who worked on that redesign: the data simply said so.

The team began working on the interface redesign in January 2011. They called the project "Density," because the new design's goal was literally a denser user experience (Figure 3.7).

The original experience had given the user four titles in a row from which to choose, with a "play" button and star rating under each title's thumbnail. Each title also had ample whitespace surrounding it—a waste of screen real estate, in the team's opinion.

The variation presented scrollable rows with title thumbnails. The designers removed the star rating and play button from the default view, and made it a hover experience instead.

They then A/B tested both variations on a small subset of new and existing members while measuring retention and engagement in both variations. The result: retention in the

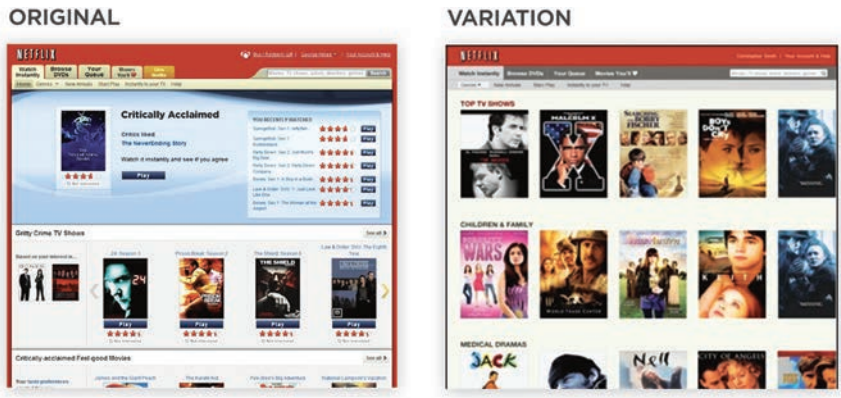

FIGURE 3.7 Netflix original site versus "Density" redesign.

variation increased by 20 to 55 basis points, and engagement grew by 30 to 140 basis points.

The data clearly told the designers that new and existing members preferred the variation to the original. Netflix counted it as a success and rolled the new "density" interface out to 100 percent of its users in June 2011. As Gumm asserts, "If [the results hadn't been] positive, we wouldn't have rolled it out." The company measured engagement and retention again in the rollout as a gut-check. Sure enough, the results of the second test concurred with the first that users watched more movies and TV shows with the new interface.

Then the comment backlash started.

However, as far as Netflix is concerned, the metrics reflecting data from existing and new members tell the absolute truth. As Gumm explains, the vocal minority made up a small fraction of the user base and they voiced an opinion that went against all the data Netflix had about the experience. Gumm points out, "We were looking at the metrics and people were watching more, they liked it better, and they were more engaged in the service. . . . [Both the tests] proved it."

Gumm also makes the following very important point: "What people say and what they do are rarely the same. We're not going to tailor the product experience, just like we're not going to have 80 million different engineering paths, just to please half a percent of the people. It's just not worth the support required."

Gumm then reminds us that despite the few loud, unhappy customers that may emerge, the most critical thing to remember is the data: "I think it's really important in an A/B testing organization, or any data-driven organization, to just hold true to the philosophy that the *data is what matters*."

TL;DR

- Incrementalism can lead to local maxima. Be willing to **explore** to find the big wins before testing smaller changes and tweaks.
- Conversely, sometimes it's the incremental refinements that prove or disprove your hypotheses about what your users respond to. **Use the insights from small tests** to guide and inform your thinking about bigger changes.
- Consider entirely new **alternative approaches** to your principal business goals. Be willing to go beyond just testing "variations on a theme"—you might be surprised.
- If you're working on a major site redesign or overhaul, don't wait until the new design is live to A/B test it. **A/B test the redesign** itself.