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

CHAPTER

6

Fail Fast and Learn

Learning to Embrace the Times When A Beats B

If you want to increase your success rate, double your failure rate.
Thomas J. Watson, Former Chairman and CEO of IBM

After what we have been showing you, it may be tempting to think that *every* experiment you run will have a positive outcome. (Why, then, even run the experiment? Just keep changing the thing that you were going to test, and it'll be great!) The reality, of course, is that not every variation is going to beat the control.

Even "failed" experiments have their silver linings, however: recognizing that a particular change will harm your goals is inarguably better than simply making that change, and as an added benefit experiments like this are often the ones that teach us the most about our visitors and what drives them. Indeed, experiments that fail tend to contradict some assumption held by the tester, and the results can point to reasons why that assumption is wrong.

Prime Real Estate versus Familiar Real Estate: IGN

Gaming website IGN wanted to encourage more visitors to the video site that brings them a big portion of their ad revenue. So they tried running an A/B test where they moved the "Videos" link over to the left of the main navigation (Figure 6.1).

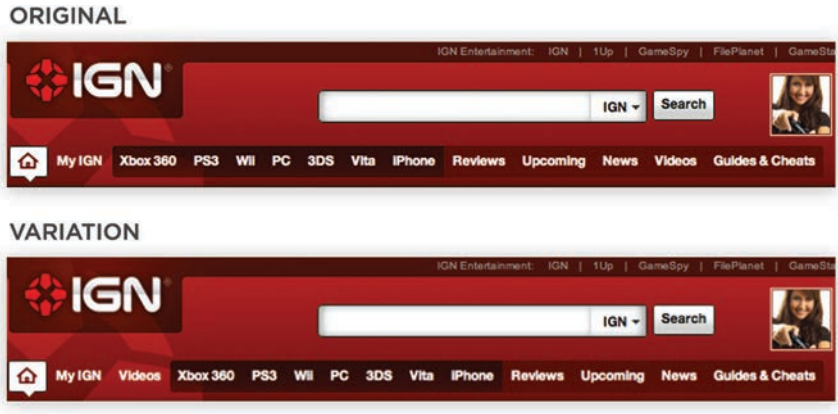

FIGURE 6.1 Original IGN navigation versus variation with repositioned "Videos" button.

There are plenty of organizations in which a change like this would have come down the chain of command once-and-for-all from the HiPPO—the Highest Paid Person's Opinion. But before making the change for good, IGN ran an experiment to see *exactly* how much of an increase they could expect to see from giving the "Video" link top billing.

Not only did the test, shockingly, show no *increase* at all, it showed that the new banner *dramatically reduced* the video click rate by 92.3 percent. If they had blindly moved the "Videos" link without testing it first, the change could have been disastrous. Because IGN gets so much traffic, it only took them a matter of hours to get statistically significant results. They were able to cease the experiment, return to the original design, and go back to the data for more answers.

The test saved IGN from a potential catastrophe that would have occurred had they simply rolled out the new navigation, but there's a bigger lesson. One of the biggest reasons for dramatic results like this one is that a lot of a site's traffic typically comes from *returning* visitors, users who are

accustomed to seeing the site in a certain way—in this case, with the "Videos" link on the far-right side, not the far-left. When it's missing from the spot they normally go to find it, they're not going to do the work to locate it. Considering the root cause of the results offers lessons not only about proposed changes but, at a deeper level, about the testing *process* as well. Moving forward, the team can consider the fact that new and returning users are going to have very different experiences of the site. Keeping this in mind will bear fruit in subsequent tests. (We take a deeper look at the insights to be gained from traffic segmentation in Chapter 13.)

What's Good at One Scale Isn't Always Good at Another: E-Commerce

A/B testing is very prominent among e-commerce businesses. Many of those we've worked with are constantly testing something on their sites. As they say, "There's always something to be improved on or optimized."

Many retailers have customer reviews and star ratings displayed on their sites. One large online retailer discovered through testing that displaying the rating prominently on *individual* product pages helped conversion. So the e-commerce team there experimented with adding the ratings to the category page—the page one level up from the product page that shows all of the items in a category. It seemed like common sense to the e-commerce team: showing the stars on the spill page should motivate people to click through and view the products more often thereby increasing conversions. (Figure 6.2.)

Good thing they tested it because that was not the case, the variation did more harm than good. As it happened, showing

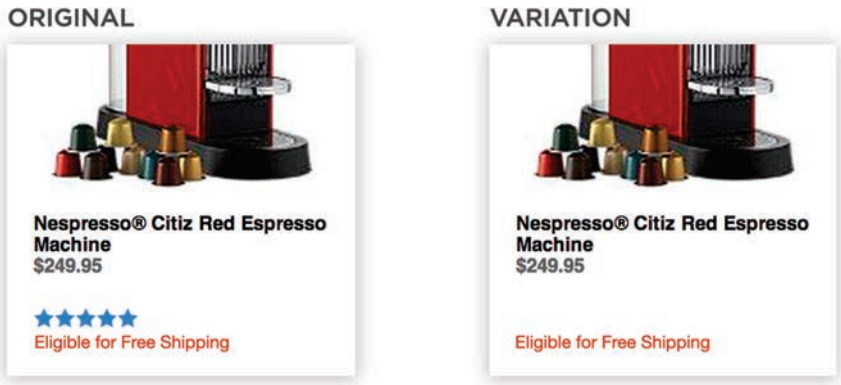

FIGURE 6.2 The original category product description with star ratings versus a variation without.

star reviews made customers convert 10 percent less. The test illustrates that *what works at one level of scale doesn't always work at another level*; what was good for the product page ended up being bad for the category page. It's a good reminder that just because something makes sense on a particular part of your site—or is even *proved* to be advantageous—doesn't mean you should roll it out to other parts of the site without checking first.

What Buyers Want Isn't Always What Sellers Want: Etsy

Handmade- and vintage-goods marketplace Etsy has over 42 million unique visitors per month and is among the Alexa Top 200 websites. A/B testing is an important means for the product developers and engineers to collect behavioral data about how Etsy's 800,000 sellers and 20 million members use the site.

Etsy users are shown an activity feed in which they can see highlights from fellow Etsy members they follow: the items those

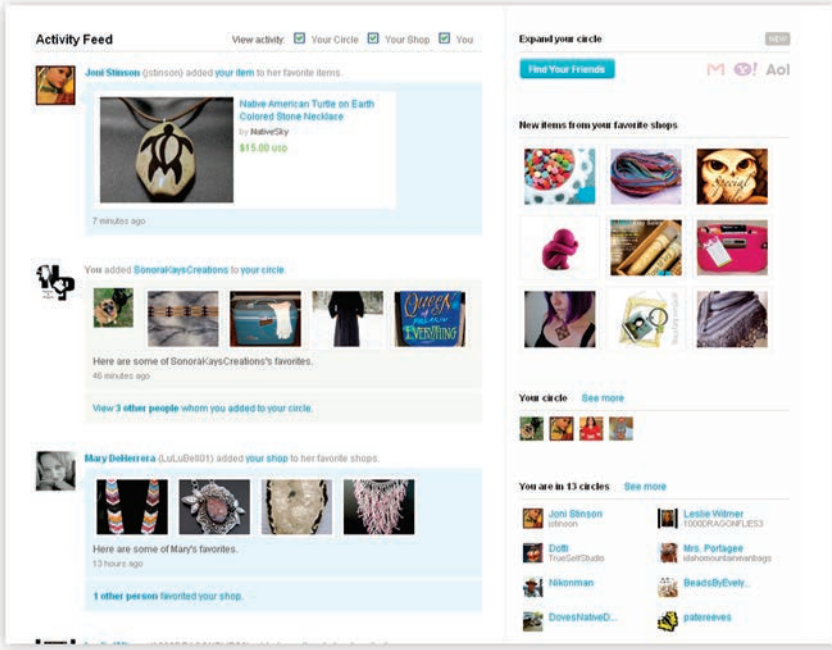

FIGURE 6.3 Original Etsy activity feed.

people are favoriting and purchasing. There are thousands of items posted to Etsy weekly, and this activity feed is a handy way for users to discover new items on the site (Figure 6.3).

The activity feed displays a combination of activities happening for buyers and sellers in one list. In what the team thought would be a much improved experience, they redesigned the feed and removed the "Your Shop" view from it, leaving only the "People You Follow" view. They A/B tested the original feed against the redesigned one to see how the redesign fared with users.

To the team's surprise, engagement with the feed dramatically decreased in the variation. After a closer look at the data, they discovered a certain type of use case that the team didn't

anticipate. It turned out that sellers were using their own activity feed to manage their shops: as a timeline of what items they had listed at what times. The team envisioned the feed as a tool for buyers to scroll through what people were doing on the site. But *sellers* had been using this to manage their shops and the new "re-skinning" removed this functionality for them.

Without this surprise result, the Etsy team would never have known about this use case. Now, not only could they take it into account during the redesign, they could actually *design for it*.

Their next iteration included two buttons: one called "Following," and another for "Your Shop" (Figure 6.4). The story ends happily with Etsy now actively building site functionality around a usage that, until their original redesign hit a snag, they had never known about.

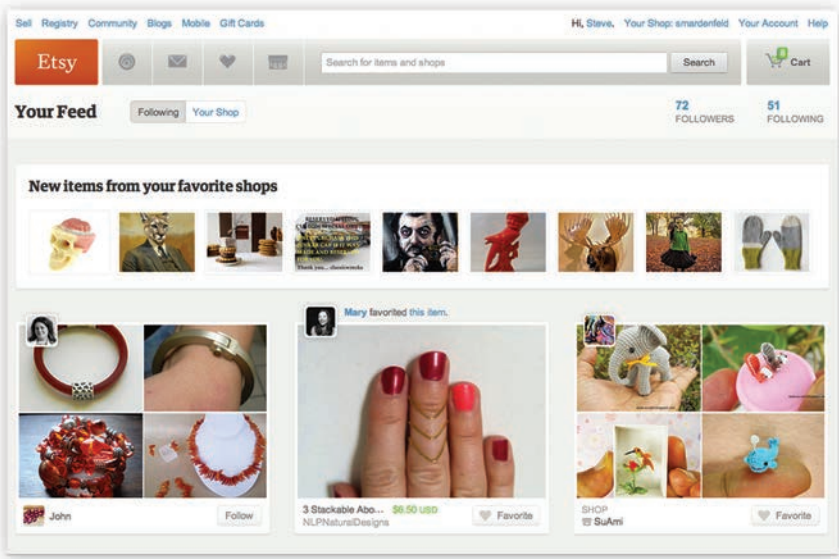

FIGURE 6.4 Final Etsy activity feed.

When a Win Isn't a Win (Is a Win): Chrome Industries

The Chrome e-commerce team has experimented with a plethora of image treatments for their urban biking products over the years, and recently decided to test whether a product *video* spurred more visitors to make purchases than did a static image (Figure 6.5).

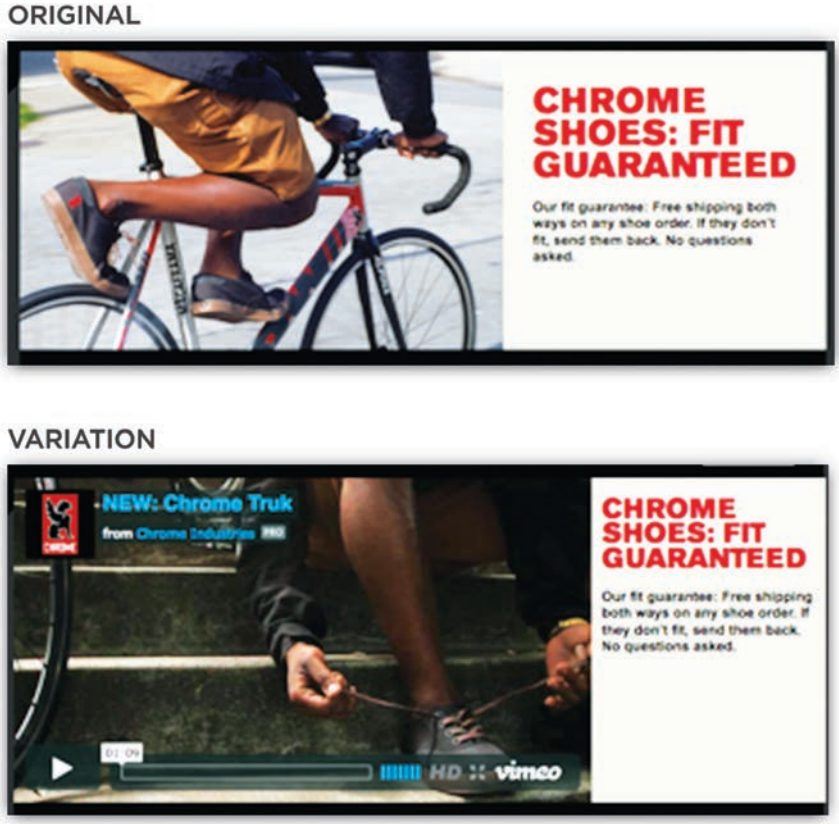

FIGURE 6.5 Original Truk shoe screenshot with static image versus variation Truk shoe screenshot with video.

The objective of the test was to determine whether to commit more resources toward video development. The team picked one product to experiment with: their Truk shoe.

They measured the percentage of visitors to the Truk product page from the category page, the percentage of visitors who continued to checkout, and percentage of visitors who successfully ordered. After letting the test run for just under three months, the results were something of a wash. Users visited the Truk product page 0.5 percent more with the image, continued to checkout 0.3 percent more with the video, and successfully ordered 0.2 percent more with the video.

If anything, the video slightly edged out the static image, but because producing video involves a much higher investment from Chrome than the images, the verdict is actually a clear vote *against* the added production cost.

Chrome can table the issue for the time being, or it can further investigate the reason why video didn't convert, rather than moving forward under the assumption that video will drive sales and ramping up a full-blown video asset initiative that won't necessarily prove its return on investment. If the team does choose to test video down the line (e.g., in seeing how lifestyle-oriented video might compare against product-oriented video), they can at least be confident that there's little risk in running the follow-up test, since they've proven that video won't hurt conversion. They may also decide simply to allocate their energies elsewhere and experiment with optimizing different portions of the site entirely, where there may be bigger unrealized gains awaiting.

TL;DR

- What works for **returning users** may not work for new users, and vice-versa.
- **Something that works on one page may not work on another**; something that works at one scale may not work at another.
- What one type of user wants may not be what another type of user wants. A failed test, sometimes more than a successful test, may prompt a drill-down that reveals a **key difference between segments of users**.
- Sometimes a variation may win against the original, but it may not **win by enough of a margin** to justify the implementation overhead or other drawbacks of the variation that are external to the test itself.
- Any test that reveals that an initiative isn't performing is a blessing in disguise: it allows you to **free up resources** from things that aren't working and divert them to the things that are.