DENIS ANDRASEC, MARCUS BLOICE, GEORG LEXER, JÜRGEN ZERNIG

A/B TESTING LITERATURE REVIEW

Denis Andrasec, Marcus Bloice, Georg Lexer, Jürgen Zernig: *A/B Testing Literature Review*, Information Architecture and Web Usability, © December 6, 2011

ABSTRACT

This report outlines and describes a data-driven design method known as A/B Testing. It covers aspects such as its history, its usage by several high-profile corporations, and demonstrates how A/B Testing can be used to test websites using frameworks such as the Google Website Optimizer.

CONTENTS

LIST OF FIGURES

1 INTRODUCTION

This chapter serves as an introduction to the A/B Testing method, and will also describe the structure of this report. In order to fully describe A/B Testing, four main topics were identified and each chapter deals with one of these topics.

In section 2 the history of A/B Testing is discussed. As a technique, A/B Testing has actually existed long before the web was born and has predecessors in several fields, including medicine. Several of these are outline in this section of the report.

In section 3, the topic of *Design vs. Data* is presented. Because A/B Testing belongs to the field of data-driven design, it has been criticized for quantifying design and subjugating creative innovation. Detractors claim that design cannot, or should not, be approached in a such a purely scientific manner, and this chapter attempts to offer an objective viewpoint of the situation.

Section 4 concentrates on highlighting a number of examples of A/B Testing in use. Many household-name companies have used A/B Testing to much success, and a number of them will be described in this section.

Last, in section 5, various A/B Testing frameworks are presented. Due to the popularity of A/B Testing, and its relative simplicity, many frameworks have appeared allowing for the easy integration of A/B tests into websites. A number of these frameworks will be compared, and their respective advantages and disadvantages discussed.

Therefore, the remainder of this section will briefly outline what A/B Testing is, how it is commonly used, and why this method has been embraced by so many companies.

1.1 WHAT IS A/B TESTING?

Put simply, A/B Testing is a method used to gain insight into visitor behavior by creating two (or more) versions of a webpage and analyzing which version results in a higher conversion rate. That means, for example, that a newly designed "Buy Me" button can be empirically tested to see whether a new color results in more sales [Smashing Magazine, 2010a].

To do this, users who visit a website are randomly placed into different groups, where each group sees a different version of the same website. In its simplest sense, two versions of a website are created and 50% of its visitors are randomly assigned to the original design (the *control* group) and 50% are assigned to the new design (the *test* or *treatment* group). A new design can be radically different, or it can be a tiny incremental change that is barely noticeable to the average user. A metric is chosen as a measure of success, and the users' behavior is then logged and later analyzed. Typically, a metric might be the percentage of users who clicked on a particular button, or who completed a signup process.

The experiment runs simultaneously; that means all versions of your design are live at the same time. Users that have been assigned to the various versions of your site should therefore never be moved to another group during a test. Normally, a cookie or something similar is used to enforce this. Figure 1 describes the basic principle of A/B Testing graphically.

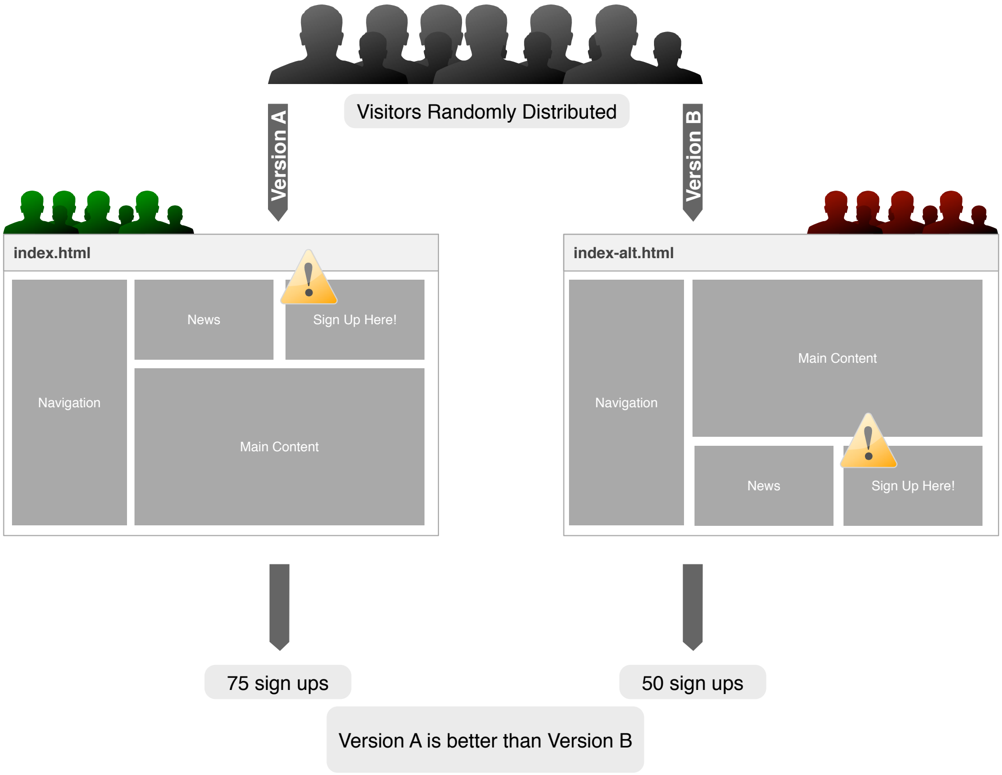


Figure 1: The above depicts an example of a simple A/B test. As can be seen, the users are split into two groups where each group sees a slightly different version of a webpage (in this case, the signup section has been moved from being above the main content to just below it). The chosen metric in this case is the number of signups. In this example, Version A's design is better than Version B's as more people signed up by the end of the test period.

The very strength of A/B Testing is its simplicity—in terms of both the ease of which it can be incorporated into a website, and the ease in which the data can be analyzed. This, however, brings up the issue of statistics and how the analyze the data correctly. As this is beyond the scope of the introductory chapter, statistical significance and data analysis are described in more detail in section 3 on page 9. However, the fact remains that A/B Testing's main advantage is its simplicity: even the most minuscule design changes can quickly and easily be tested before becoming part of a website's newer design—with little or no disruption to the service or the user's experience, and therein lies its greatest asset.

2 HISTORY

As already mentioned in the introduction, there were several testing techniques before A/B Testing came on the horizon and they were used long before the internet. It is important to take a look at them to understand the basic value of testing your designs, versus just going with your gut and relying on "known" truths.

This chapter will take a look at three different examples of testing before A/B Testing on the web was employed. The first one is a book about testing, Scientific Advertising from Claude Hopkins. Hopkins describes the value of testing versus long discussion in his work. The second one is a case study from how Toyota deploys testing to optimize their processes and factories. Thirdly, this chapter shows how medical trials are structured and why they are similar to testing. After these examples, A/B Testing in the context of web is examined. It is discussed why testing should be used and how the criterion for evaluating results is constructed. Finally the results of this historical look at testing are examined and shown what is important for testing.

2.1 TESTING IN THE TIME OF ADVERTISING DISASTER.

In his famous book Scientific Advertising Claude Hopkins says: "Almost any questions can be answered, cheaply, quickly and finally, by a test campaign. And that's the way to answer them - not by arguments around a table. Go to the court of last resort - the buyers of your product." (1923)

When looking at advertising, there are many factors that influence the potential success or failure of an advertising campaign. That is because tastes differ so much, that a person alone can hardly predict what the majority of people will like. One person may be sure of a good idea for an ad campaign and it fails, when on the other side people may laugh at an idea at first, but it becomes a huge success. Exactly for this reasons, this period was called the time of advertising disaster. Because advertisers ventured on their own ideas, huge amounts of money got spend on big campaigns. Only a few guessed right, and most of them guessed wrong and lost a lot of money. It wasn't known to them how much their cost per customer or profit per customer was [Hopkins, 1923].

So what should be done? Hopkins suggested that small ventures could be made in the form of testing campaigns and let the thousands decide what the millions want. If the data is known for a few thousand people and their buying habits for a specific product/design, from that the data can be interpolated and applied to more people. So big risks can be avoided at minimal cost, because the outcome of an idea can be more or less calculated based on data from a small controlled test group [Hopkins, 1923].

Hopkins went on to demonstrate this way of thinking with an example. A large food advertiser wanted to try a new design for one of their products.

Most of the involved parties were convinced that the product would be more popular in the new form and they believed so much in it, that they wanted to try it without involving the consumer but for their luck, a test campaign was done. An ad with a coupon for the new product was dealt to customers in a few towns, and they would later be called per telephone and asked how they liked it. Almost all of them disapproved of the new design. Another design was proposed and tested, but the advertising parties didn't think it was worthwhile. From the few thousand people that got asked how they liked it, 92% gave a positive answer [Hopkins, 1923].

This shows, that with testing a lot of money can be made and also saved. Reduce risks by running tests with small controlled groups. In a lot of cases the consumer knows exactly what he wants and his opinions should be considered before making big decisions or changes to products.

2.2 PLACEBO VS TREATMENT. TESTING IN MEDICAL TRIALS.

In medical trials were always used for development, safety control and efficiency testing of drugs. They are a way of to collect data for these types of health interventions.

The subjects (plants, animals, people) that participate in such trials are divided in a control group (control) and a test group (treatment). By comparing the results of the two groups, the efficiency of the treatment is examined.

Following designs are applied to avoid biases and false data:

PLACEBO CONTROLLED
The subjects in the control group get a placebo, i.e. a drug that has no effect.

RANDOMIZED
Each subject either receives either the treatment or the placebo at random.

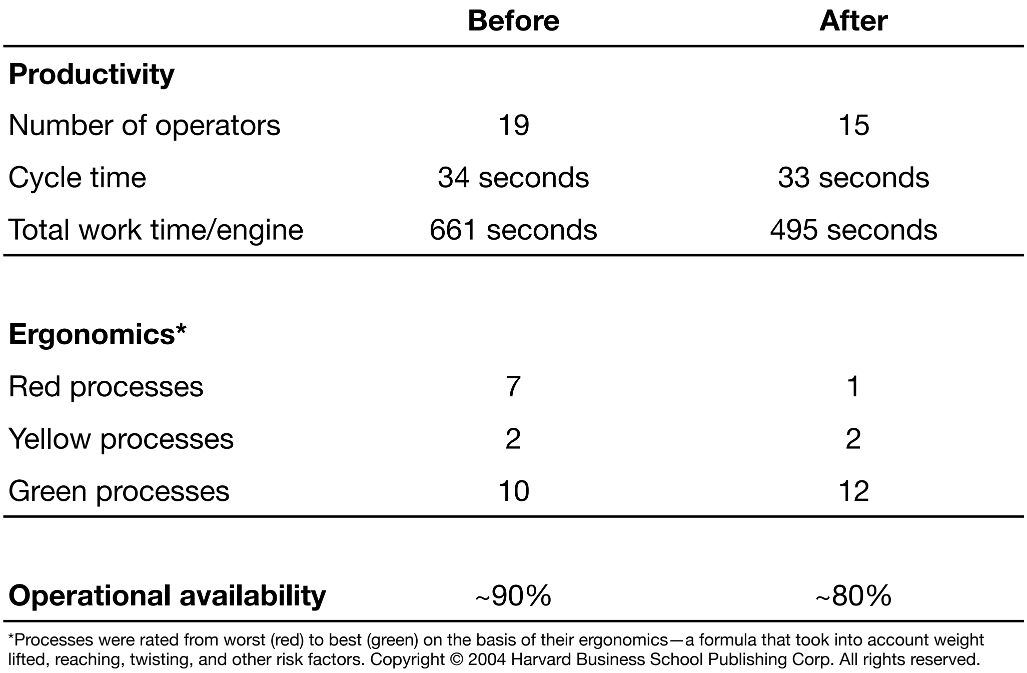
DOUBLE BLIND
Neither the patient knows if they are getting the treatment or placebo, nor the doctor who is monitoring the trial and treating the patient.

2.3 TESTING AT TOYOTA.

The people at Toyota have made testing and optimizing a core principle of the way they work. With their production system they have thrived in all areas including quality, reliability, productivity, cost reduction and other areas [Spear, 2004].

As an example, an american toyota manager had the task to improve the operation of an US engine plant. Over a course of 6 weeks, 25 changes were made. From rearranging racks, placement of parts and also placement of machine controls. All of the changes were done as experiments. The manager at first observed how something worked. Then he needed to think of how to change something and would the desired outcome was. This new configuration was then tested and the desired outcome was compared with the real outcome.

This was also the data that got presented to his superior executive. At Toyota they really wanted to understand why something worked better, not just that it worked better [Spear, 2004]. The results from these reconfigurations and tests are shown in figure 2 below.



	Before	After
Productivity		
Number of operators	19	15
Cycle time	34 seconds	33 seconds
Total work time/engine	661 seconds	495 seconds
Ergonomics*		
Red processes	7	1
Yellow processes	2	2
Green processes	10	12
Operational availability	~90%	~80%

*Processes were rated from worst (red) to best (green) on the basis of their ergonomics—a formula that took into account weight lifted, reaching, twisting, and other risk factors. Copyright © 2004 Harvard Business School Publishing Corp. All rights reserved.

Figure 2: Progress at Toyota engine plant. Significant improvements with 25 changes over the course of 6 weeks.

There are a some important lessons that can be learned from the way testing is done at toyota:

1. There is no substitution for observation.
2. Proposed changes should be tried out as experiments.
   a) Make an assumption about the problem.
   b) What are the expected results?
   c) Test.
   d) Compare expected results with real results.
3. Do small experiments quick and frequently.

2.4 A/B TESTING AND THE WEB.

A/B Testing is the process of comparing two designs and let users decide which one is the better by a criterion which is within the bounds of statistical confidence. This will be explained in detail in a later chapter, but basically there are two or more groups of designs that get compared to each other and evaluated as shown in the figure 3 below.

History has shown that testing has always been an essential process to get better results. Not only was testing used to sell more products, and develop

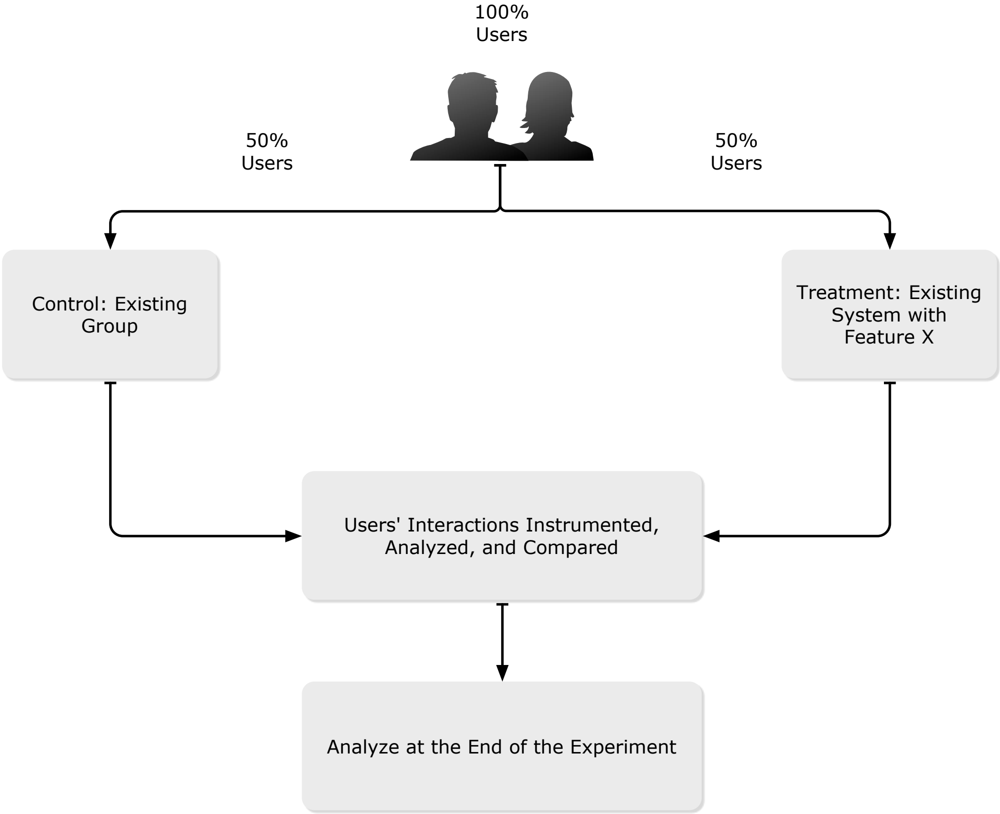

100%
Users
50%
Users
50%
Users
Control: Existing Group
Treatment: Existing System with Feature X
Users' Interactions Instrumented, Analyzed, and Compared
Analyze at the End of the Experiment

Figure 3: A/B Testing Flowchart. Users are split into two groups, control and treatment. Data is collected and later analyzed with an OEC

better drugs, but also to optimize whole product processes. So now with the web it is possible to rapidly test new designs and better products in an fast and iterative way with minimal costs and risk. There is this great opportunity to let the users decide which design/product is best by just letting them try it out and then make a decision with the newly collected data.

But tho make such an decision, there needs to be a quantitive measure on how to evaluate the outcome of such tests or experiments [Crook et al., 2009].This measure is called the OEC (Overall Evaluation Criterion) and it ultimately decides which version is the best. When choosing an OEC there are several things that should be considered but it can be broken down in one sentence: Consider the long term effects, not just the short term effects. For example if a new buy button is tested, you shouldn't just consider the click-through rate but also other factors like time on site, visit frequency and if the user is buying more than one product.

Another important tool is the ramping up of the things you want to test. If for example two designs have to be tested, you want to show one design to 50% of your users and the other one to the other 50% of your users. But it should be avoided to do this at once. Instead you should slowly increase the percentage of users who the new design is shown to. Starting at 0.1%, then 1%, then 5% and so on until 50% of the user base are reached. This is important so the experiment can be aborted if the treatment's OEC is significantly worse than the control [Kohavi, 2007].

2.5 WHAT HISTORY TEACHES US ABOUT TESTING.

History shows us that testing is a very important part of almost all processes and can improve most of them. It reduces risks and money spent. Below are the the things that a lot of the before mentioned examples have in common and what they tech us about how to tackle the task of testing:

DATA TRUMPS INTUITION
: Data should always be trusted more that intuitions and opinions. If there is a way to test a theory or idea, it should almost always be done.

REPLACE OPINIONS WITH AN OEC
: If a test was executed, the results should be evaluated and quantified with a strong OEC and not just with opinions or simple measures.

COMPUTE THE STATISTICS CAREFULLY
: It should be a principle to always try to eliminate biases and wrong data.

EXPERIMENT OFTEN AND ITERATE
: Changes should be small and incremental. Even if a big change would be a good idea, it's always safer to do small changes due to unpredictable behavior of big ones. With small changes rapid iteration is possible.

3 DESIGN VS. DATA

Plato is my friend – Aristotle is my friend – but my greatest friend is truth.[1]
—Sir Isaac Newton, Quaestiones Quaedam Philosophicae, c. 1664 [Wikiquote, 2011].

This chapter discusses the issue of Design Vs. Data, or more specifically the data-driven design paradigm and its usefulness and applicability in the design of websites. Data-driven design, also known as data-driven development, is paradigm which stipulates that design decisions should be made on empirical data and not on personal beliefs or biases. It states that all design choices, no matter how small, should be backed up by data, statistics, and testing.

Due to this very scientific approach, data-driven design, and in turn A/B Testing, have been criticized for removing, or at least reducing, the creative element of design.

This section will discuss the advantages of using A/B Testing to try to improve a website's design, while highlighting its limitations and situations where data-driven design may not suitable.

3.1 WHY USE A/B TESTING

A/B Testing certainly allows the designer to avoid certain pitfalls when making changes to a website's design. The following is an inconclusive list of issues that can detrimentally affect design choices which A/B Testing can help mitigate against [20bits, 2008]:

EVERYONE IS BIASED
: All designers have certain biases that affect designs they create. Some of these biases may be accurate, others not. A/B Testing ensures only those designs that genuinely bring an improvement are actually implemented.

THERE ARE NO "SACRED COWS"
: There are many design paradigms which are considered untouchable and almost infallible or sacred. These design guidelines are occasionally made with very little or no data to back up their claims. A/B tests can empirically test these claims—simply allow the data to drive your design choices. Do not rely on generic design guidelines that were made long before your website was even conceived.

UNIVERSAL BELIEFS ARE THE MOST DANGEROUS
: Those beliefs regarding design that are considered "givens" are the most dangerous of all, as the designer can believe these to be true without even realizing it or questioning why. Back up design choices with data and expel bias.

[1] Original: Amicus Plato – amicus Aristoteles – magis amica veritas.

ARE YOU FAILING OR SUCCEEDING? Without data, it can be difficult to tell if you are failing or even if you are *succeeding*. It may be possible that a webmaster is pleased with 100 user registrations per month—but without analyzing the data, they have no way of knowing if this is indeed a good number or percentage. It is possible that a silly design error is causing a large percentage of users to never complete the registration process, thus resulting in many lost registrations or customers.

DATA TRUMPS INTUITION To summarize: believe in the data that you have gathered in order to be sure of any design changes to your website. As Kohavi is often quoted as saying, “Data Trumps Intuition” [Kohavi et al., 2007].

Another advantage to using A/B Testing is its simplicity. By using frameworks, such as the Google Website Optimizer, it is almost trivial to test every aspect of a design against data. It may be worthwhile to do so simply for the fact that the web offers the first possibility to perform such control/test experiments in such an easy way. With little or no disruption to the user, A/B Testing makes it simple to justify design changes and alterations.

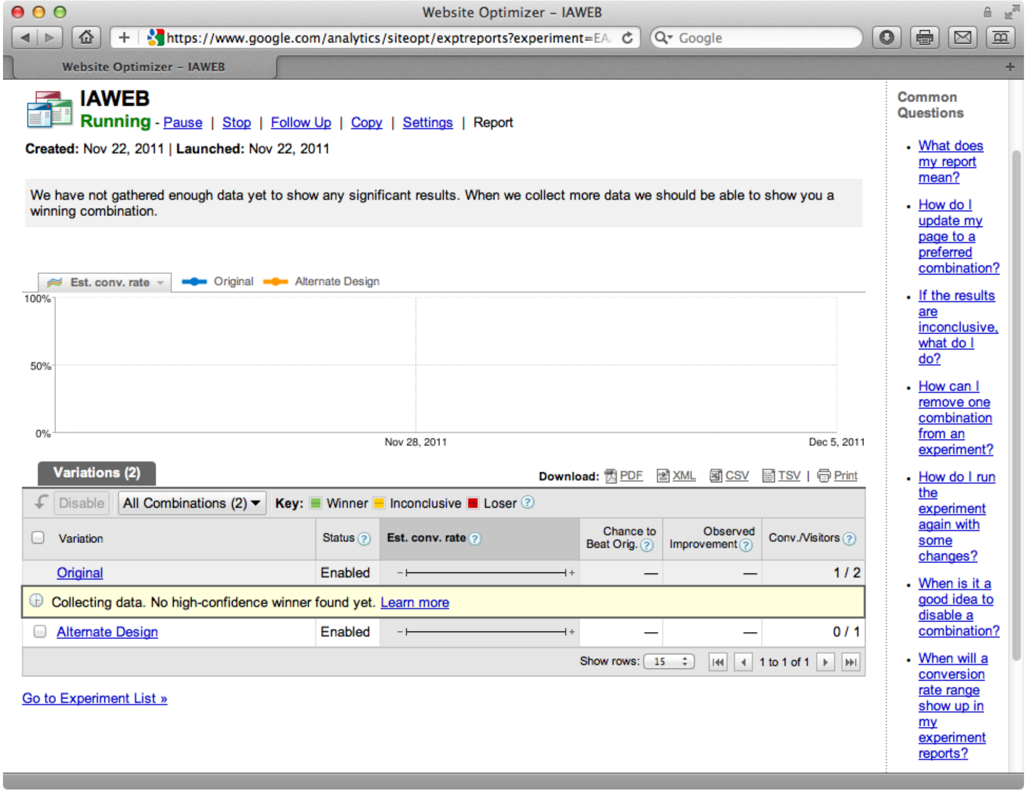
3.2 STATISTICAL SIGNIFICANCE

One crucial aspect when using A/B Testing within a website is to ensure that a design change’s impact on a chosen metric is statistically significant. In other words, that the numbers point to a high-confidence winner of a design with enough confidence so as to make it reasonable to choose this design choice over another. Figure 4 shows the Google Website Optimizer in use[2]. As can be seen, the software will not report on a winner as neither design has a more statistically significant number of conversions.

Because most frameworks (discussed in detail in section 5) calculate whether a design change is statistically significant, it is often not necessary to calculate this manually. However, calculating the confidence intervals manually is not difficult and can be estimated using the following formula: $\mathrm{acc}_h \pm z \cdot \mathrm{StdDev}$, where acc_h is the estimated mean (or click-through rate, conversion) and z is set to 1.96 for a confidence interval of 95% [Kohavi, 1995, 2007].

By ensuring that data is statistically significant, the designer ensures that natural fluctuations in visitor demographies do not influence the experiment. For example, consider a global website that serves Asia as well as Europe and North America. In Asia, color symbolism is quite different from that in the Western Hemisphere and running an A/B experiment over a short time period while visitors from China peak in numbers would result in very different results than if ran over the course of a few weeks. In China red symbolizes good luck, prosperity and happiness, while in the West it generally symbolizes danger, anger, or even importance (a red carpet) [Smashing Magazine, 2010b].

[2] http://www.google.com/websiteoptimizer


Website Optimizer – IAWEB
https://www.google.com/analytics/siteopt/exptreports?experiment=EA
Google
Website Optimizer – IAWEB
Est. conv. rate
Original
Alternate Design
100%
50%
0%
Nov 28, 2011
Dec 5, 2011

IAWEB

Running - Pause | Stop | Follow Up | Copy | Settings | Report

Created: Nov 22, 2011 | **Launched:** Nov 22, 2011

We have not gathered enough data yet to show any significant results. When we collect more data we should be able to show you a winning combination.

Variations (2)

Download: PDF XML CSV TSV | Print

Disable | All Combinations (2) ▾ | **Key:** 🟩 Winner 🟨 Inconclusive 🟥 Loser

□ Variation	Status	**Est. conv. rate**	Chance to Beat Orig.	Observed Improvement	Conv./Visitors
Original	Enabled	- ⊢———⊣ +	—	—	1 / 2
Collecting data. No high-confidence winner found yet. Learn more					
□ Alternate Design	Enabled	- ⊢———⊣ +	—	—	0 / 1

Show rows: 15 | 1 to 1 of 1

Go to Experiment List »

Common Questions

- What does my report mean?
- How do I update my page to a preferred combination?
- If the results are inconclusive, what do I do?
- How can I remove one combination from an experiment?
- How do I run the experiment again with some changes?
- When is it a good idea to disable a combination?
- When will a conversion rate range show up in my experiment reports?

Figure 4: The Google Website Optimizer will not report on a design choice before a high confidence winner is found.

For this reason also, it is recommended to run tests for all long as necessary, until a clear winner can be seen. Either use an A/B Testing framework to calculate statistical significance for you, or make these calculations manually.

3.3 Limitations

So far, the advantages of A/B Testing have been discussed and it is clear that it is a useful tool for fine-tuning a website's design, or for deciding upon two or more design choices such as the placement of a graphic or section of a website. However, it has its limitations that one must bear in mind:

- A/B Testing cannot on its own make a good design, and can only decide which design option is better. All tested designs could be very bad indeed, and A/B Testing will only help decide which is the better of a bad bunch.
- It is considered that if every decision is made on data, daring decisions will not be made [Porter, 2009]. As a startup, you may want to differentiate from the status quo, and it may well be in your interests to try to stand out or be noticed. Using A/B Testing to fine tune every last element of your site is possibly not be the best approach to making a design that gets a website noticed.

It is very important for anyone about to implemented A/B Testing in their website that they understand exactly what A/B Testing is meant to achieve. In reality, A/B Testing is more about making small incremental changes to a design rather than an attempt to replace the creative process. This can be seen in the following diagram [Porter, 2011]:

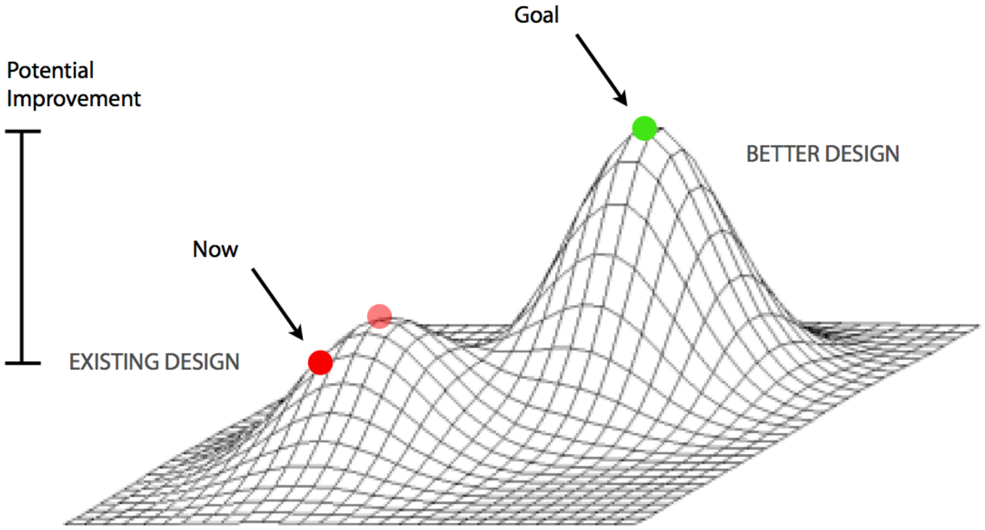


Figure 5: Incremental changes are possible with data-driven design, but design still requires creative input in order to begin with.

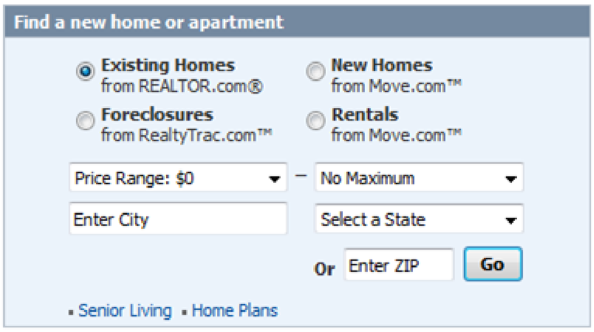
Figure 5 shows that incremental changes can improve design, but that optimization can only go so far. Designers still need to make bold jumps to reach the next level of innovative design. What is clear is that a combination of the two methods is optimum—designers must use their expertise to create interesting websites, and engineers must tackle the problem of improving these designs using methods such as A/B Testing. Finding the correct balance is key to creating an experience that users are both happy to use and find aesthetically pleasing.

4 EXAMPLES

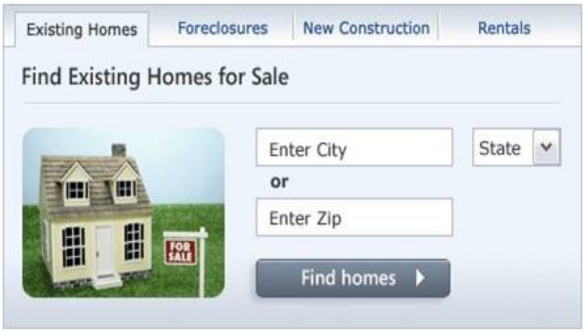
This chapter shows some examples and results of A/B Testing in use.

4.1 EXAMPLE 1: MSN REAL ESTATE

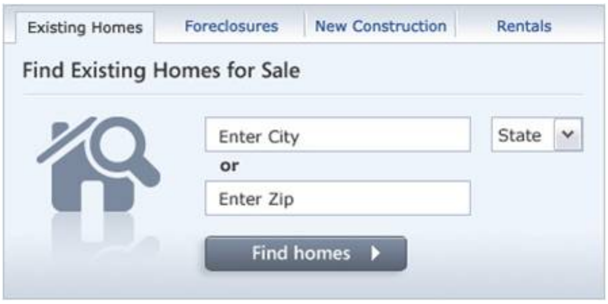
The first example is MSN Real Estate. The Overall Evaluation Criterion is in this case is the revenue to Microsoft generated every time a user clicks. They tested following the "Find a house" widget variations:


Find a new home or apartment
● Existing Homes from REALTOR.com®
○ New Homes from Move.com™
○ Foreclosures from RealtyTrac.com™
○ Rentals from Move.com™
Price Range: $0
–
No Maximum
Enter City
Select a State
Or
Enter ZIP
Go
Senior Living
Home Plans

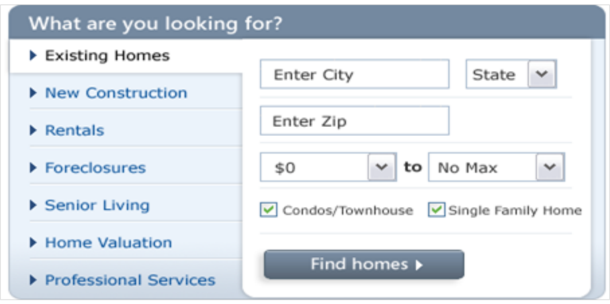
Variation 1



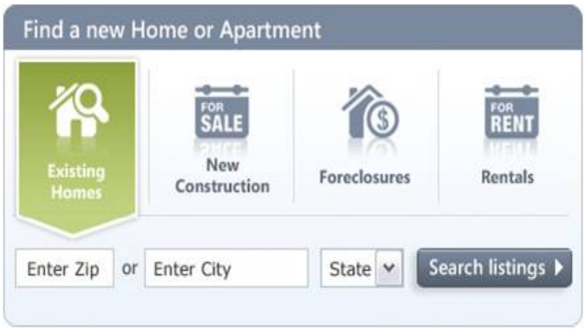
Variation 2



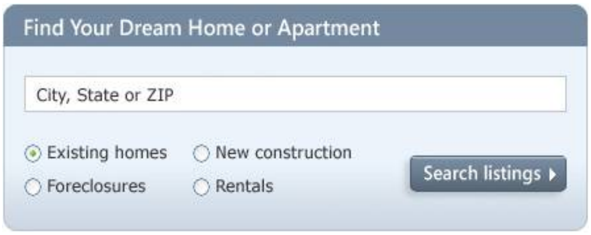
Variation 3



Variation 4



Variation 5

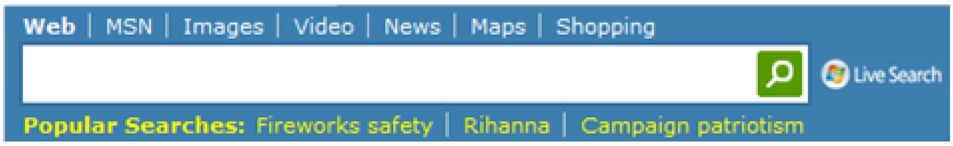


Variation 6

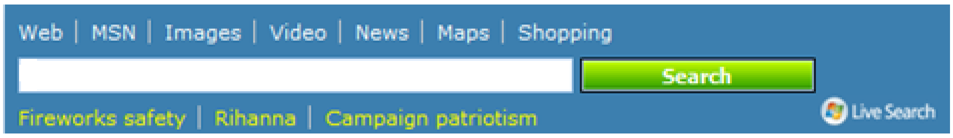
The widget which performed best was against every expectation and turned out to be Variation 6. The principle of keep it small and simple should never be underrated. The revenue increased with this variation by about 9.7% [Kohavi, 2009].

4.2 EXAMPLE 2: MSN HOME PAGE SEARCH BOX

In this example the Overall Evaluation Criterion was the click-through rate for the search box and the popular searches links.


Web | MSN | Images | Video | News | Maps | Shopping
Live Search
Popular Searches: Fireworks safety | Rihanna | Campaign patriotism

Variation 1


Web | MSN | Images | Video | News | Maps | Shopping
Search
Fireworks safety | Rihanna | Campaign patriotism
Live Search

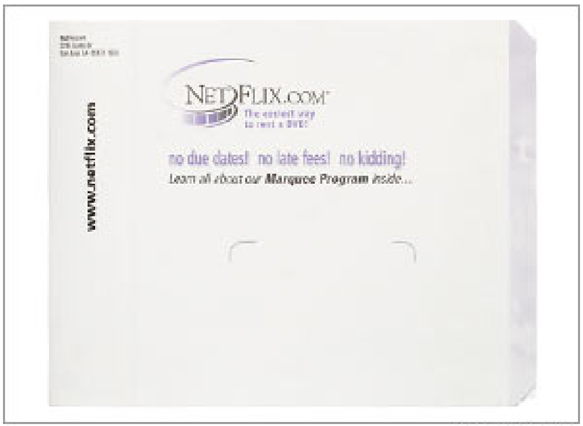
Variation 2

As can be seen, the differences are that Variation 1 has a taller search box (however, the overall size is the same) and Variation 2 has magnifying glass icon, links to some "popular searches" while Variation 2 has large search button rather than a magnifying glass.

In this case, however, no remarkable differences were found [Kohavi, 2009].

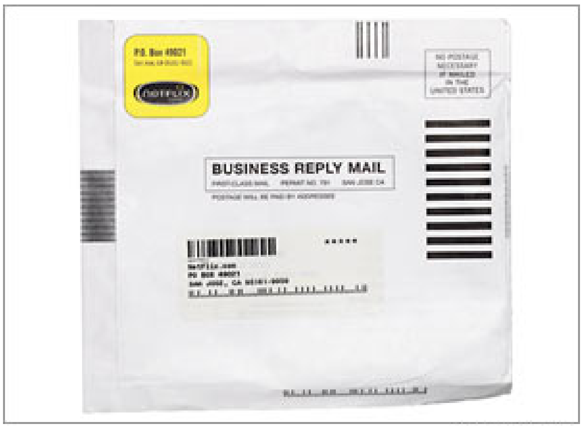
4.3 example 3: netflix mailer evolution

In 2006, CNN reported on the evolution of the Netflix mailer packaging. The results of user feedback were used to alter and iterate the design over a number of years in order to create an optimal design for the packaging. This section describes this evolution.


www.netflix.com
NETFLIX.COM
no due dates! no late fees! no kidding!
Learn all about our Marquee Program inside...

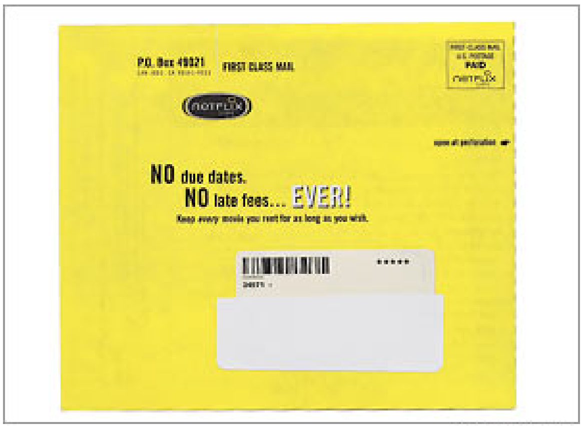
NUCCI STUDIO

This above layout is from 1999, is made from cardboard, and weighed more than an ounce. At the time Netflix only had approximately 100,000 customers and did not spend much effort thinking about how to reduce material and shipping costs.



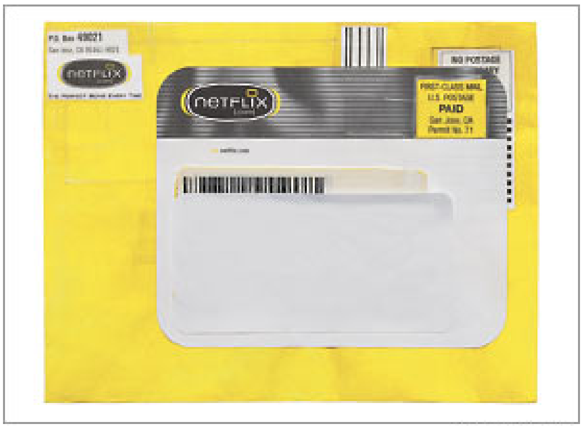
NUCCI STUDIO

This above layout is from 2000 and made from thick paper instead of cardboard. Discs are removed and inserted from the top rather than from the side.


P.O. Box 49021
FIRST CLASS MAIL
FIRST-CLASS MAIL U.S. POSTAGE PAID netflix
netflix
NO due dates.
NO late fees... EVER!
Keep every movie you rent for as long as you wish.

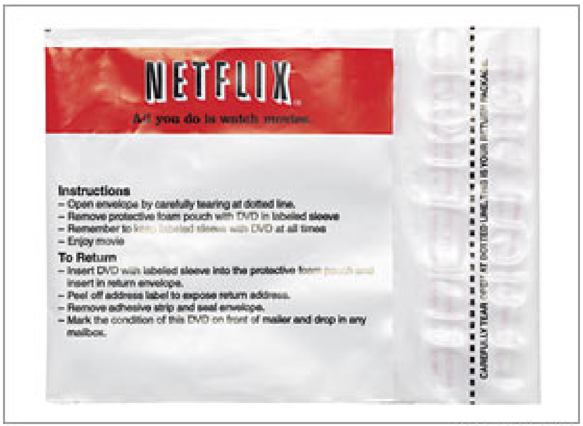
NUCCI STUDIO

In the same year, they introduced color printing. They changed to side-loading, which was meant to be more convenient.



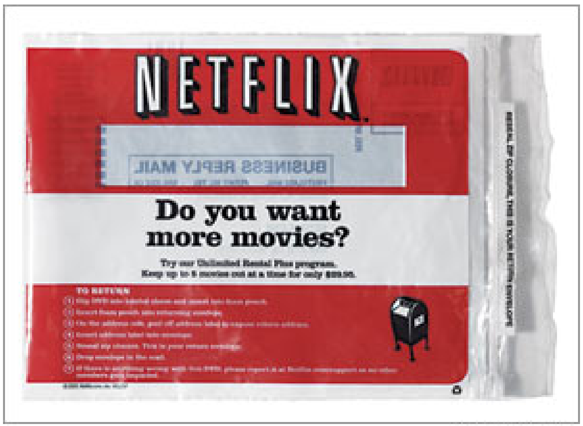
NUCCI STUDIO

The next change occurred in the same year: customers had to peel off a sticker to get the return address.


NETFLIX
All you do is watch movies.
Instructions
– Open envelope by carefully tearing at dotted line.
– Remove protective foam pouch with DVD in labeled sleeve
– Remember to keep labeled sleeve with DVD at all times
– Enjoy movie
To Return
– Insert DVD with labeled sleeve into the protective foam pouch and insert in return envelope.
– Peel off address label to expose return address.
– Remove adhesive strip and seal envelope.
– Mark the condition of this DVD on front of mailer and drop in any mailbox.
CAREFULLY TEAR OPEN AT DOTTED LINE. THIS IS YOUR RETURN PACKAGE.

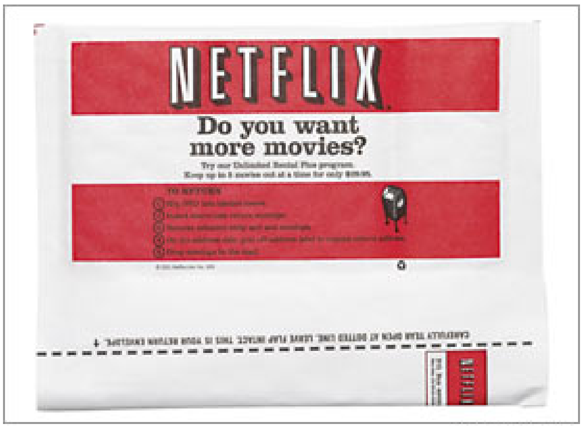
NUCCI STUDIO

They then tried changing to plastic instead of paper. Plastic is cheaper, but problems occurred when packages would inflate when being transported on airplanes.



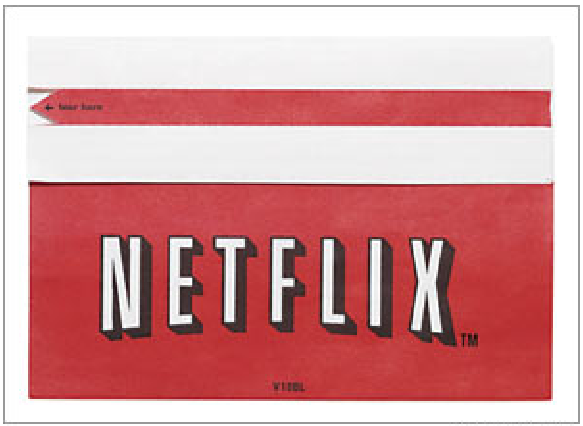
NUCCI STUDIO

In 2001 they added an air-hole to prevent inflation.



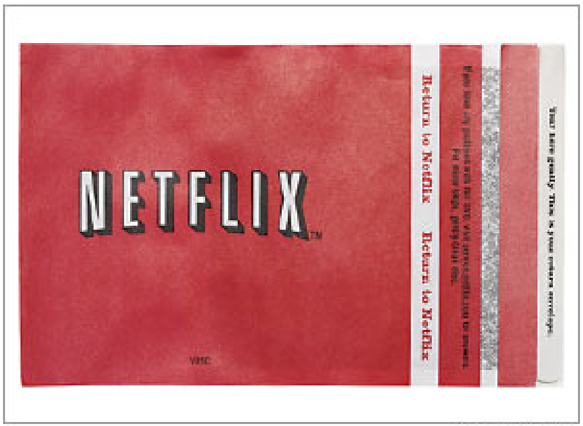
NUCCI STUDIO

In the same year they returned to paper and added foam paddings to reduce breakage.



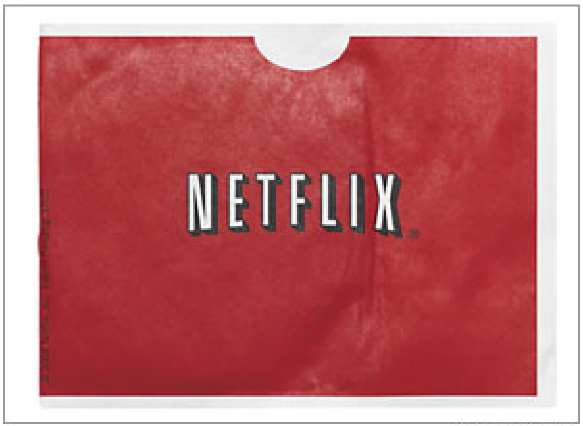
NUCCI STUDIO

Some weeks later they dropped the padding due to costs. They once again changed to a top-loading mechanism.



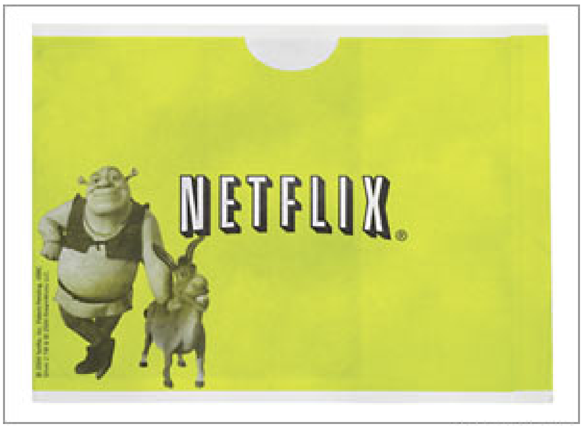
NUCCI STUDIO

Again, side-loading was reintroduced.



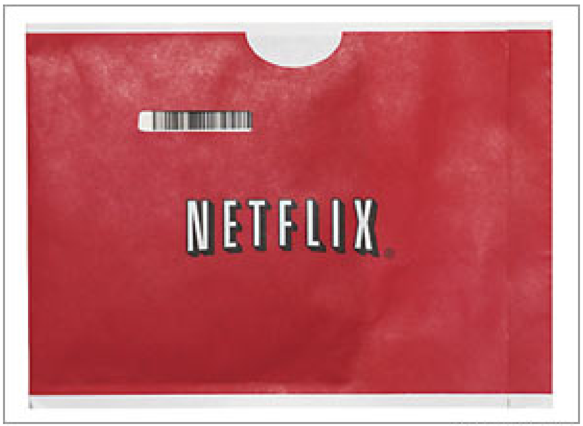
NUCCI STUDIO

In 2003 they introduced a circular sticker, bonded to the top.



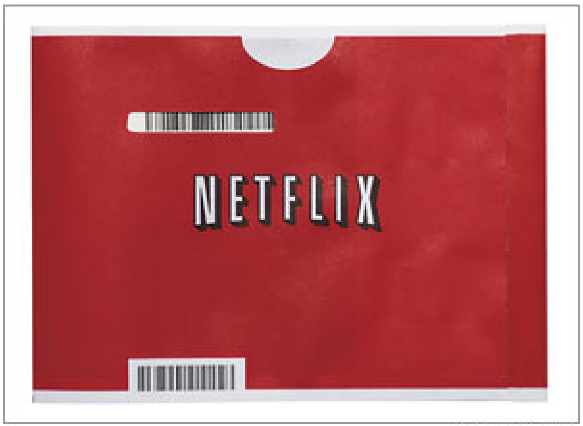
NUCCI STUDIO

In 2004, Netflix started using the mailer to promote the release of DVDs.



NUCCI STUDIO

In the same year, they added a window that shows the disc barcode.



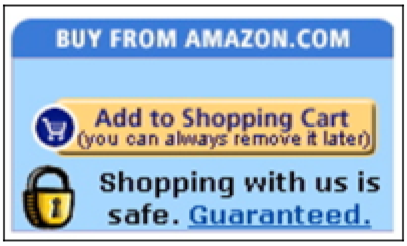
NUCCI STUDIO

In 2005 Netflix added one more barcode at the bottom of the envelope. This is the current mailer and the result of five years of experimentation.

See [CNN, 2006].

4.4 example 4: amazon's add-to-cart evolution

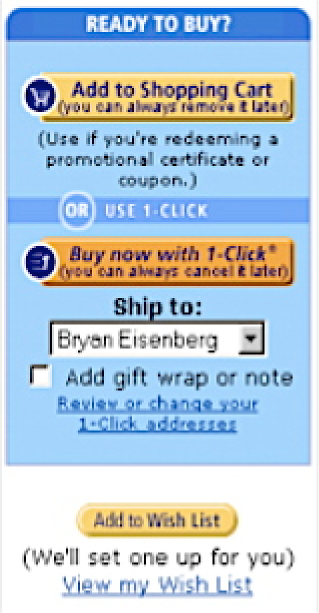
This example shows the Evolution of the Add-to-Cart button from amazon. The overall evaluation criterion is the number of orders. Amazon are one of the evangelists of A/B Testing and have used it extensively in their data-driven design approach.



Shopping with us is safe. Guaranteed.

The above is an early version of the button. The "(you can always remove it later)" and the "Shopping with us is safe. Guaranteed." text can be seen. Customers were anxious of the button in Amazon's early years as they thought that they would instantly buy an item as soon as they clicked the button.

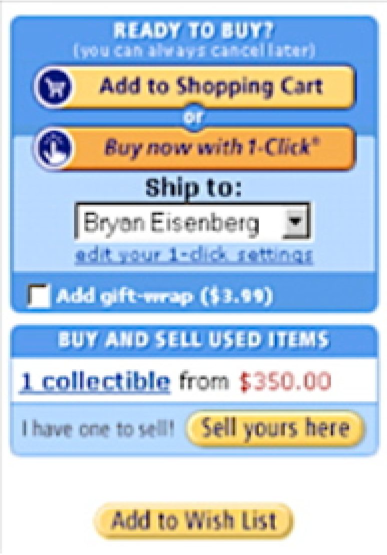
Therefore, Amazon decided to make the *Add-to-Cart* button less ambiguous for its customers.



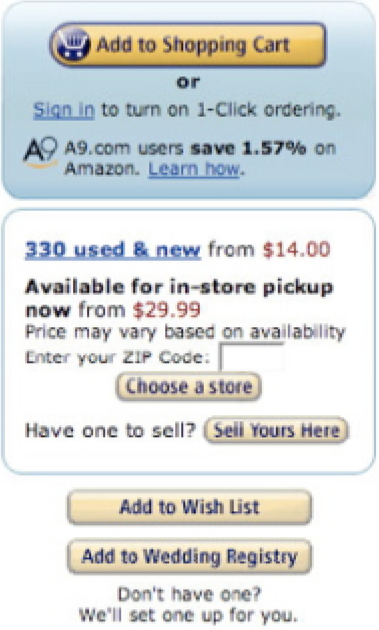
(Use if you're redeeming a promotional certificate or coupon.)

(We'll set one up for you)

This was done by replacing the “Buy from Amazon.com” with “Ready to Buy?” and introduced the “Buy now with 1-Click*” button. This is also the first time the “Add to Wish List” button appears.



They redesigned the layout and made two stand-alone areas with a noticeable partition of both areas. They removed the “you can always remove it later” from the Add-to-Cart button and placed it up to the “Ready to Buy?” area and changed the text to “you can always cancel later”.



Sign in to turn on 1-Click ordering.

A9.com users **save 1.57%** on Amazon. Learn how.

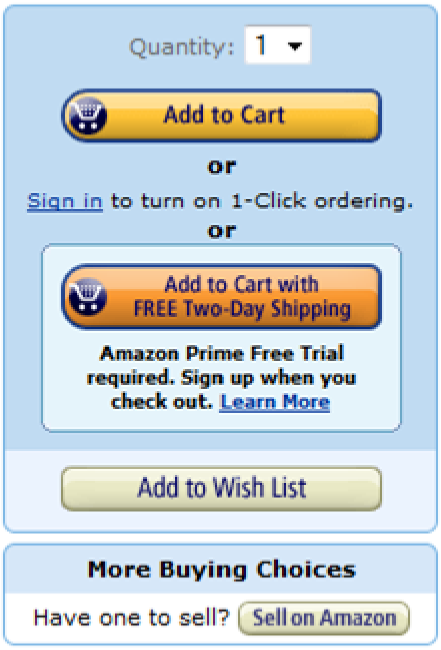
330 used & new from $14.00

Available for in-store pickup now from $29.99
Price may vary based on availability
Enter your ZIP Code:

Have one to sell?

Don't have one?
We'll set one up for you.

The Amazon 2.0 version now has a 3D effect. The "Ready to Buy?" text does not exist anymore. The "Buy with 1-Click" button appears after login and they started promoting the A9 search engine.



Quantity: 1

Sign in to turn on 1-Click ordering.

Amazon Prime Free Trial required. Sign up when you check out. Learn More

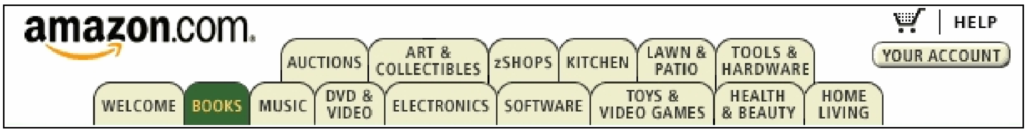
More Buying Choices

Have one to sell?

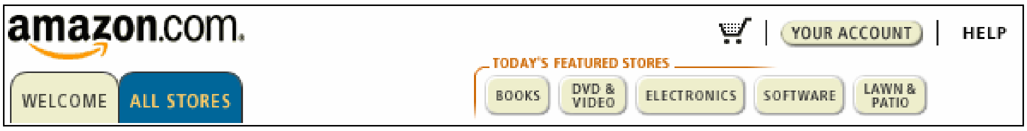
Above is the current version of the button. They have added a pull-down menu to adjust the quantity and stopped promoting the A9 search engine [Grok Dot Com, 2008].

4.5 example 5: amazon's tabs

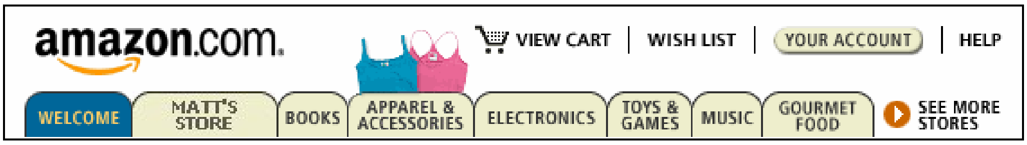
Another short example is Amazon's usage of tabs.



Over time, more and more tabs appeared for each of their departments, which resulted in a reduced amount of clicks on to the tabs. This was because the location of the tabs would change too frequently.



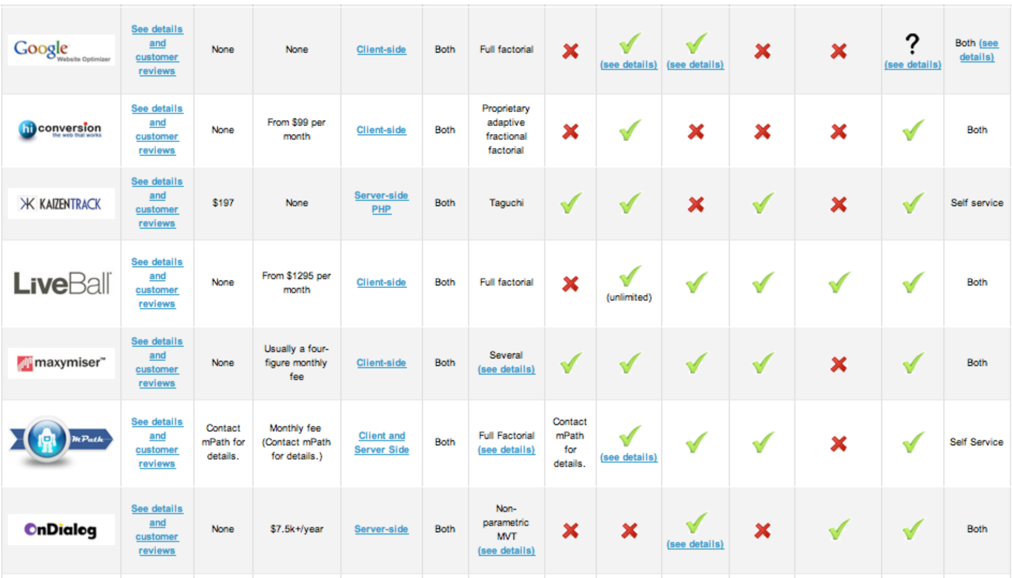
To counter this, Amazon decided to delete most of the buttons. However this version proved to be relatively minimalistic. Thorough testing ensued to determine the optimal number of tabs.



The end result was a balance between keeping perspective and covering the main departments and areas of the store.

5 FRAMEWORKS

This chapter introduces a number of frameworks that can be used to implement A/B tests into a live website in order to test design iterations. A very good comparison of many multi-variant tools can be found at http://www.whichmvt.com, a screenshot of which can be seen in figure 6.



Google Website Optimizer	See details and customer reviews	None	None	Client-side	Both	Full factorial	✗	✓ (see details)	✓ (see details)	✗	✗	? (see details)	Both (see details)
hi conversion the web that works	See details and customer reviews	None	From $99 per month	Client-side	Both	Proprietary adaptive fractional factorial	✗	✓	✗	✗	✗	✓	Both
KAIZENTRACK	See details and customer reviews	$197	None	Server-side PHP	Both	Taguchi	✓	✓	✗	✓	✗	✓	Self service
LiveBall	See details and customer reviews	None	From $1295 per month	Client-side	Both	Full factorial	✗	✓ (unlimited)	✓	✓	✓	✓	Both
maxymiser™	See details and customer reviews	None	Usually a four-figure monthly fee	Client-side	Both	Several (see details)	✓	✓	✓	✓	✗	✓	Both
mPath	See details and customer reviews	Contact mPath for details.	Monthly fee (Contact mPath for details.)	Client and Server Side	Both	Full Factorial (see details)	Contact mPath for details.	✓ (see details)	✓	✓	✗	✓	Self Service
OnDialog	See details and customer reviews	None	$7.5k+/year	Server-side	Both	Non-parametric MVT (see details)	✗	✗	✓ (see details)	✗	✓	✓	Both

Figure 6: Screenshot of the Which Multivariate website

Other tools are shown and also compared in the table shown in figure 7 in section 5.1.

- Google Website Optimizer[1]
- A/Bingo[2]
- Visual Website Optimizer[3]
- Kaitzentrack[4]
- Reedge[5]

5.1 COMPARISON OF SELECTED FRAMEWORKS

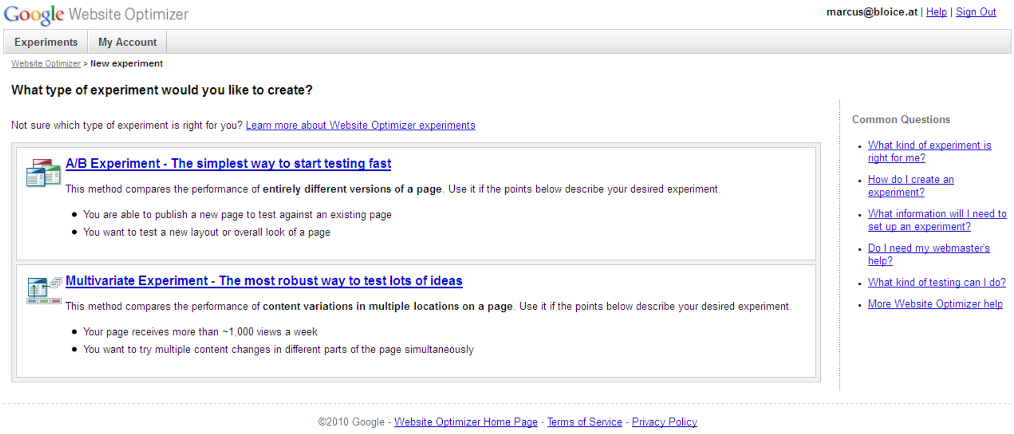
[1] http://www.google.com/websiteoptimizer/ Accessed: 05 Dec 2011
[2] http://www.bingocardcreator.com/abingo Accessed: 05 Dec 2011
[3] http://visualwebsiteoptimizer.com/ Accessed: 05 Dec 2011
[4] http://www.kaizentrack.com/. Accessed: 05 Dec 2011
[5] http://www.reedge.com/ Accessed: 05 Dec 2011

	A/B Test	Multivariant Test	Type of Multivariant	Technology	Multiple Goals	Pricing	Support
Google Website Optimizer	X	X	full factorial	client-side	X	Free	Self
A/Bingo	X	X	full factorial	server-side (RoR)	X	Free	Self
Visual Website Optimizer	X	X	full factorial	client-side	X	$49	Self
Kaitzentrack	X	X	Toguchi	server-side (PHP)	X	$179	Self
Reedge	X	X	full factorial	client-side	X	$29/m	Self

Figure 7: Framework Comparison

5.2 GOOGLE WEBSITE OPTIMIZER - SCREENSHOTS

Some Screenshots of Google Website Optimizer[6] to show how the A/B Testing works in general.



Website Optimizer > New experiment

What type of experiment would you like to create?

Not sure which type of experiment is right for you? Learn more about Website Optimizer experiments

A/B Experiment - The simplest way to start testing fast

This method compares the performance of **entirely different versions of a page**. Use it if the points below describe your desired experiment.

- You are able to publish a new page to test against an existing page
- You want to test a new layout or overall look of a page

Multivariate Experiment - The most robust way to test lots of ideas

This method compares the performance of **content variations in multiple locations on a page**. Use it if the points below describe your desired experiment.

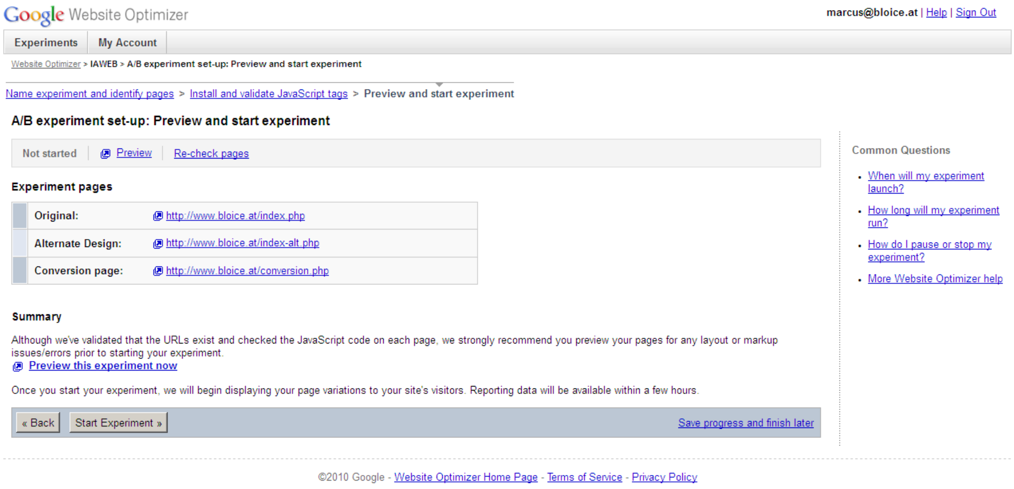
- Your page receives more than ~1,000 views a week
- You want to try multiple content changes in different parts of the page simultaneously

Common Questions

- What kind of experiment is right for me?
- How do I create an experiment?
- What information will I need to set up an experiment?
- Do I need my webmaster's help?
- What kind of testing can I do?
- More Website Optimizer help

©2010 Google - Website Optimizer Home Page - Terms of Service - Privacy Policy

Figure 8: Google A/B Testing



Website Optimizer > IAWEB > A/B experiment set-up: Preview and start experiment

Name experiment and identify pages > Install and validate JavaScript tags > Preview and start experiment

A/B experiment set-up: Preview and start experiment

Experiment pages

Original:	http://www.bloice.at/index.php
Alternate Design:	http://www.bloice.at/index-alt.php
Conversion page:	http://www.bloice.at/conversion.php

Summary

Although we've validated that the URLs exist and checked the JavaScript code on each page, we strongly recommend you preview your pages for any layout or markup issues/errors prior to starting your experiment.
Preview this experiment now

Once you start your experiment, we will begin displaying your page variations to your site's visitors. Reporting data will be available within a few hours.

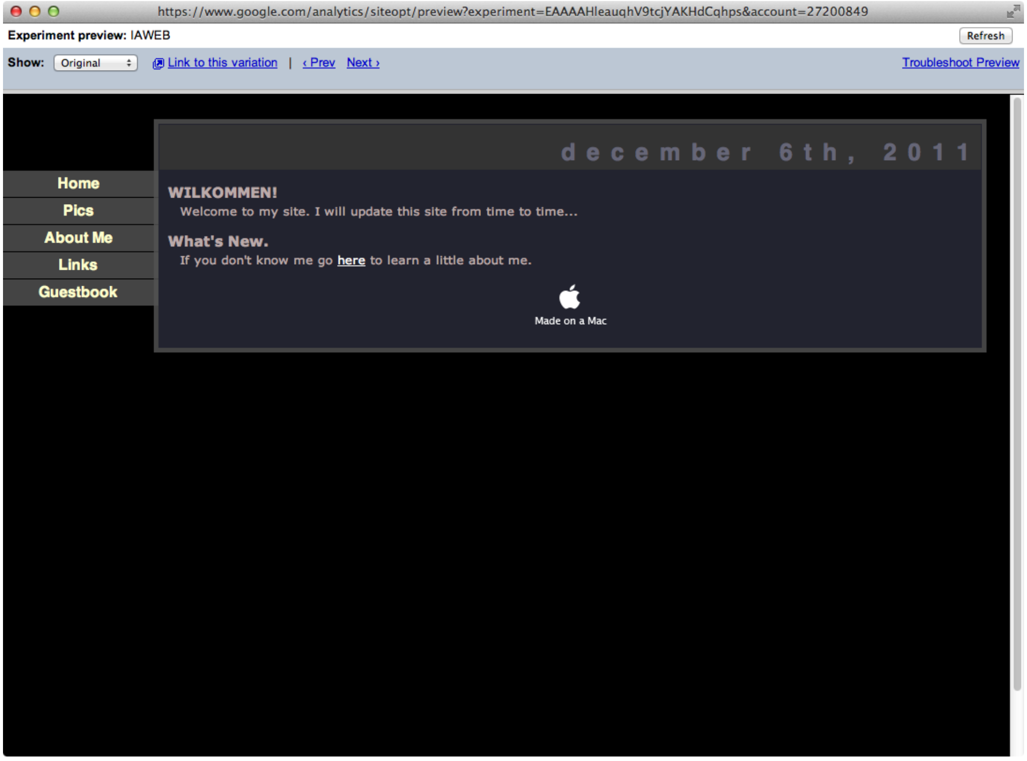
Common Questions

- When will my experiment launch?
- How long will my experiment run?
- How do I pause or stop my experiment?
- More Website Optimizer help

©2010 Google - Website Optimizer Home Page - Terms of Service - Privacy Policy

Figure 9: Google A/B Testing Set-up

[6] http://www.google.com/websiteoptimizer/ Accessed: 05 Dec 2011



https://www.google.com/analytics/siteopt/preview?experiment=EAAAAHleauqhV9tcjYAKHdCqhps&account=27200849

Experiment preview: IAWEB

Show: Original

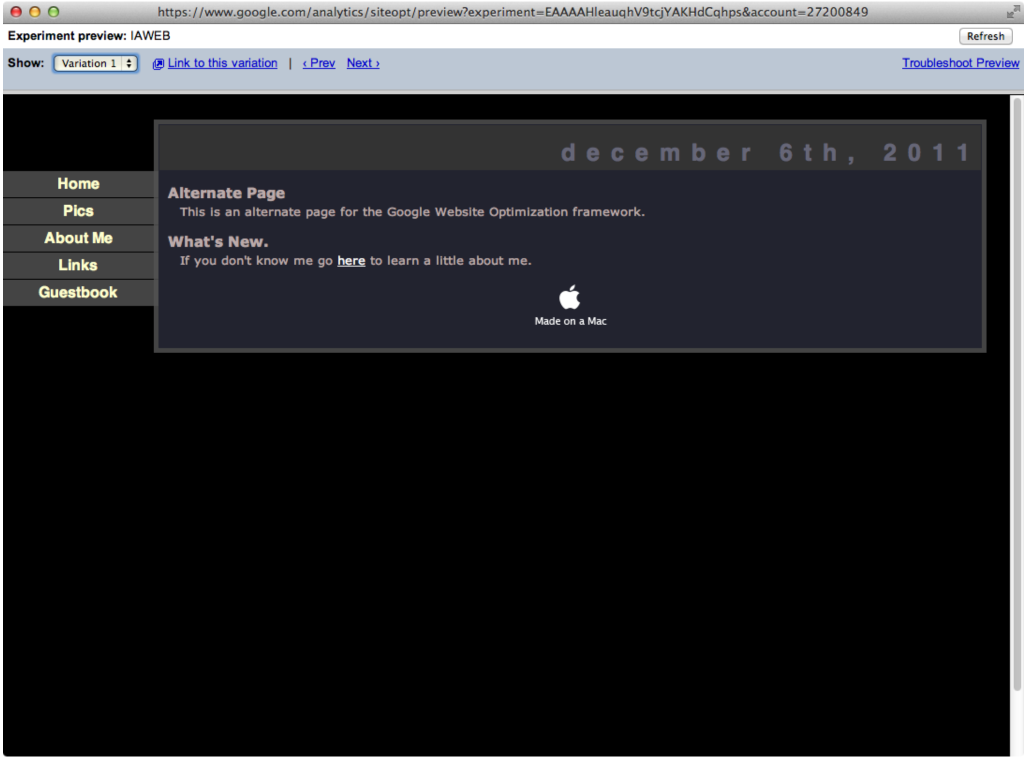
WILKOMMEN!

Welcome to my site. I will update this site from time to time...

What's New.

If you don't know me go here to learn a little about me.

Figure 10: Original design - done with Google Website Optimizer



https://www.google.com/analytics/siteopt/preview?experiment=EAAAAHleauqhV9tcjYAKHdCqhps&account=27200849

Experiment preview: IAWEB

Show: Variation 1

Alternate Page

This is an alternate page for the Google Website Optimization framework.

What's New.

If you don't know me go here to learn a little about me.

Figure 11: Variation 1 - done with Google Website Optimizer

BIBLIOGRAPHY

20bits. Data-driven development. http://20bits.com/articles/data-driven-development/, 2008. Accessed: 06 Dec 2011.

CNN. Netflix Mailers. http://money.cnn.com/popups/2006/biz2/netflix/frameset.exclude.html, 2006. Accessed: 05 Dec 2011.

Thomas Crook, Brian Frasca, Ron Kohavi, and Roger Longbotham. Seven Pitfalls to Avoid when Running Controlled Experiments on the Web. *KKD*, 2009.

Grok Dot Com. Hidden Secrets of the Amazon Shopping Cart. http://www.grokdotcom.com/2008/02/26/amazon-shopping-cart/, 2008. Accessed: 05 Dec 2011.

Claude C. Hopkins. *Scientific Advertising*. Cosimo Inc, 1923.

R. Kohavi, R.M. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 959–967. ACM, 2007.

Ronny Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *International joint Conference on Artificial Intelligence*, volume 14, pages 1137–1145, 1995.

Ronny Kohavi. Practical Guide to Controlled Experiments on the Web. Microsoft Presentation, 2007.

Ronny Kohavi. Planning, Running, and Analyzing Controlled Experiments on the Web. Tutorial Presentation, http://robotics.stanford.edu/~ronnyk/2009-06-28KDDTutorialT4part1.pdf, 2009.

Joshua Porter. Goodbye, Google. http://stopdesign.com/archive/2009/03/20/goodbye-google.html, 2009. Accessed: 05 Dec 2011.

Joshua Porter. Metrics-Driven Design. SXSW Presentation, 2011.

Smashing Magazine. The Ultimate Guide to A/B Testing. http://www.smashingmagazine.com/2010/06/24/the-ultimate-guide-to-a-b-testing/, 2010a. Accessed: 05 Dec 2011.

Smashing Magazine. Color Theory for Designers. http://www.smashingmagazine.com/2010/01/28/color-theory-for-designers-part-1-the-meaning-of-color/, 2010b. Accessed: 05 Dec 2011.

Steven J. Spear. Learning to Lead at Toyota. *Harvard Business Review*, 2004.

Wikiquote. The Ultimate Guide to A/B Testing. http://en.wikiquote.org/wiki/Isaac_Newton, 2011. Accessed: 05 Dec 2011.