

## Bad Data

The 3 validity threats that make your tests look conclusive (when they are deeply flawed)

### ABSTRACT

In this Web clinic transcript, Dr. Flint McLaughlin explained the three validity threats that many marketers overlook when running tests, triggers to identify if they are threatening your tests, and how to mitigate the risks to the data you collect.



# Bad Data

## The 3 validity threats that make your tests look conclusive (when they are deeply flawed)

### Presenters

Dr. Flint McGlaughlin  
Austin McCraw  
Phillip Porter

*[Note: This document is a transcript of our original Web clinic on [validity threats](#) that originally aired September 14, 2011.]*

### Writers

Paul Cheney  
Austin McCraw

### Editors

Daniel Burstein  
Selena Blue  
Brad Bortone

### Technical Production

Luke Thorpe  
Steven Beger  
Jessica McGraw

### Contributors

Beth Caudell

### Educational Funding provided by:



**Dr. Flint McGlaughlin:** It's hard to believe we have another clinic yet again today. These have just marched on year after year, after year. We have a lot of interesting data, and this is a clinic unusual in the sense that we're going to delve deeply into the actual testing process. The title is 'Bad Data: The 3 validity threats that make your tests look conclusive (when they're deeply flawed).'

In fact, I just want to say this as we go into the whole clinic, as it has been prepared, and the research and the briefings that we will share with you. It's been my experience, having been engaged in the net back when it was still ... well, prior to the World Wide Web and prior to HTML, that I've watched the whole notion of testing become a part of the expectations associated with an e-marketing effort. I remember in the early days when almost no one was testing. We were insisting that you need to test. And, it's been encouraging to see marketers everywhere test. And, then our latest benchmark study, released at MarketingSherpa, we saw that people and companies everywhere are engaged in optimization work and even developing new positions within their organization dedicated to optimization. Still, when I talk to optimization leaders within these companies, when I review the data sets from tests, all over the world, I see a common problem. Almost no one is truly testing against those validity threats, which can inhibit the quality of your data and thus leave you in a position you're trying to make a decision with flawed assumptions. What I'm trying to say is I think perhaps 75- to 80% of all of the test being run on the Internet right now have validity threats that are undetected.

Now, this is a significant problem, so as I start to get into the briefing with you and show you examples of what I'm talking about, I want to really sober my audience. Forgive me if I'm a bit tense or passionate about this, but there is nothing worse than making a confident decision with poor data. It's one thing ... it's almost better if you don't test at all if you can't test properly, because without the false set of or the false sense of confidence that poor data can give you, you're more careful. But, give a

marketer clear data that suggests path X is best, then that marketer has every reason to jump in with a lot of budget and a lot of effort, and drive in a direction that could be actually costing them millions and millions of dollars. I'm glad that marketers all over the Internet have awakened to the fact that testing is so significant, but I'm worried that we're doing it in sloppy ways, without proper training. Now, I want to balance that. You don't have to be a scientist to run good tests. In fact, if you're on this call today and you say, "Look, we just feel lucky we're testing," then this whole clinic is still for you, because ordinary people who are determined to do this right, without a Ph.D. in Statistics or in some related discipline, should be able to run effective, useful tests. But, you need to know some fundamentals to make that possible.

So, I, as always, will wade right into a case study and show you an example, a very recent example, but I'd just like to invite you to pay careful attention as we go forward because I think we're touching something right now that is virtually a plague. Our study at Sherpa tells us that only 40% of the people running tests are even doing the basics of sample size or confidence intervals. I mean, the reality is that core, essential aspect of testing isn't even being done. And, yet those who do have no idea about instrumentation effect, perhaps, or historic effect, or could give me a proper definition, or recognize it in a data set. So stay tuned. We're going to spend the rest of this time investing every single moment doing everything we can do to increase your understanding of how to run effective tests, where you have safe data sets that will allow you to make solid decisions.

You can use the hashtag #WebClinic throughout this presentation to communicate with your peers, and I'm joined today by two of the experts on our team. One of them is a Senior Editorial Analyst, Austin McCraw. You've probably heard Austin on other clinics. He is one of the most knowledgeable people we have here, with regards to the methodology, and he is deeply engaged right now with me on a book on value propositions. And, I am also joined by Phillip Porter, but we haven't heard from Phillip a lot. He is one of the big brains that hides in the back room and crunches data all day long. He has a small title that does not adequately capture the big job he must do. I have seen him and one of his colleagues, Bob Kemper, turn a box of numbers into a wealth of insights, and that is no small task. So, he's really a statistician, a mathematician and an expert in this area, and he's going to help us. And, so with Phillip and Austin, we move straight to a case study.

---

**Experiment: HubSpot Lead Gen Test**

---



**Experiment ID:** HubSpot Lead Gen Test

**Location:** MarketingExperiments Research Library

**Test Protocol Number:** TP3055

**Research Notes:**

**Background:** HubSpot teamed up with MarketingExperiments and the Optimization Summit audience to create a treatment email offer landing page for a free chapter of the MarketingSherpa LPO Benchmark Report

**Goal:** To increase HubSpot's knowledge of their current email list as well as to grow it

**Primary research question:** Which landing page will result in more chapter downloads?

**Approach:** A/B multifactor split test

**Dr. Flint McGlaughlin:** This was a recent study done in conjunction with our latest Optimization Summit. HubSpot teamed up with MarketingExperiments and the Optimization Summit audience to create a treatment for, basically, an email offer landing page. And, it offered a free chapter of a MarketingSherpa *LPO Benchmark Report*. The goal was to increase HubSpot's knowledge of their current email list as well as to grow it. So, we wanted more people to take the download so we could capture more names. So, we were looking for an increase, an increase in conversion rates associated with the landing page that would help us get the most possible email addresses. So, keeping that in mind, we designed the test around our classic type of question, which focuses on a 'which' factor (i.e., which landing page) would result in more chapter downloads.

## Experiment: Control



**Download the Free Chapter!** **Discover the Key Components of a Successful Landing Page Optimization Strategy**

First Name \*

Last Name \*

Email (privacy policy) \*

Phone \*

Company \*

Company Website \*

Role at Company \*  
- Please Select -

Number of Employees \*  
- Please Select -

Does Your Business Primarily Sell to Other Businesses (B2B) or Consumers (B2C)? \*  
- Please Select -

Does Your Business Provide PR, Web Design, SEO, or other Marketing Products or Services? \*  
- Please Select -

Biggest Marketing Challenge

**Download Now!**

**Free chapter from MarketingSherpa's first-ever *Landing Page Optimization Benchmark Report*.**

Landing page optimization (LPO) is becoming more affordable as it grows in sophistication. In fact the ROI for successful LPO programs is impressive.

According to MarketingSherpa, to be successful LPO must have an objective. That can be as simple as generating a lead or closing a sale. But instead of looking only at the business aspects, you should also frame your objectives in terms of human behavior. In other words, LPO is optimization of websites for specific human behaviors.

The chapter *Key Components of a Successful Landing Page Optimization Strategy* is about what your website visitors want rather than how to drive them to your website. In it you will find out the different methodologies marketers employ to:

- Determine visitor motivation
- Generate relevant experiences
- Create a model for quantitative evaluation

Based on data from 2,673 marketers, this chapter will give you quantifiable information on:

- The use and effectiveness of dedicated landing pages vs. your default website
- The value of competitive intelligence
- The kinds of metrics collected, and the effectiveness of visitor data in optimizing for relevance
- Lead quality scores and who is using it
- Marketing insights on how to create relevant and targeted pages

**Start maximizing the ROI of your website traffic today. Download your free copy of *Key Components of a Successful Landing Page Optimization Strategy* by filling out the form to your left.**

[Tweet This](#)
[Share on Facebook](#)
[Share on LinkedIn](#)



**Dr. Flint McLaughlin:** And, here is the control. I want you to watch this carefully because I'm going to be asking you for some feedback. And, with me is Austin, and I'm going to let Austin just kind of take you through the design of this experiment. I'll just tell you the challenge. How do you get 200+ marketers to agree on a treatment and to actually produce together something that's in testable format where we might get something with statistical confidence. So, go ahead, Austin.

**Austin McCraw:** Yeah, let me just up the ante there a little bit. There were a couple of more challenges to it, to throw on top of trying to get 200 marketers to agree. One of those is the fact that we were somewhat limited in our test ability here, what we could test. I mean, one of the first things we'd want to test on a page like this is the amount of form fields. But, that was the first constraint we gave the audience, the marketers. You can't change number of form fields. You're going to have to work around that.

**Dr. Flint McGlaughlin:** Yes.

**Austin McCraw:** Another big constrainer/challenge really is that, and it's almost hard to believe, but the motivation coming to this page was incredibly high. We'll show you in a minute, but the conversion rates were somewhere in the upper 40s at this point. So, what that means is there is one out of every two people coming to this page is converting. When we have motivation levels that high, it becomes very challenging to increase the performance. It's very ...

**Dr. Flint McGlaughlin:** And, even if you do, you typically see the increase in tiny increments ...

**Austin McCraw:** Yes.

**Dr. Flint McGlaughlin:** ... because again ... and just to point this out, for those of you that are familiar with us, remember  $C = 4m + 3v + 2(i-f) - 2a$ . Four, that coefficient in front of M, is closely connected to what Austin just said, incoming motivation is the highest factor in terms of your conversion rate. Keep going, Austin.

## Experiment: Variable Options

---

**Variable:** Headline

**Control:**

### Discover the Key Components of a Successful Landing Page Optimization Strategy

Free chapter from MarketingSherpa's first-ever Landing Page Optimization Benchmark Report.

**Option 1:**

### Maximize your marketing ROI with the MarketingSherpa Landing Page Optimization Benchmark Report

Free chapter from MarketingSherpa's first-ever Landing Page Optimization Benchmark Report.

**Option 2:**

### Get a FREE chapter from the 2011 MarketingSherpa Landing Page Optimization Benchmark Report (list price \$447)

Discover the Key Components of a Successful Landing Page Optimization Strategy

**Option 3:**

### Get a FREE 40-page Chapter Containing the Latest Landing Page Optimization Research from MarketingSherpa

Discover the Key Components of a Successful Landing Page Optimization Strategy

**Austin McCraw:** Yeah. So, what we did is we broke the audience down to groups of three and then we began to give them variables in which they could vote for values for those variables. What I mean by that is we break the page into the headline, right? So, you can test the headline, and here are the four options you have for that headline. And, so they would have to debate amongst themselves in the groups, and figure out which headline they thought would perform best. So, for instance, in this example, they chose Option #3.

Another variable that we tested was the call-to-action. You really don't have to get all of the options here, but we just want to kind of give you a sense of the process. But, they selected, I think, Option #1.

**Variable:** Call-to-Action

**Control:**

Download Now

Option 1:

Get Your Free Chapter Now

Option 2:

Get the Latest LPO Research

Option 3:

Read More

\*\* CTA Option #3 to go with Copy Option #3 only

**Austin McCraw:** We also tested the image. Here are the four options that we gave them. They chose Option #3.

**Variable:** Image

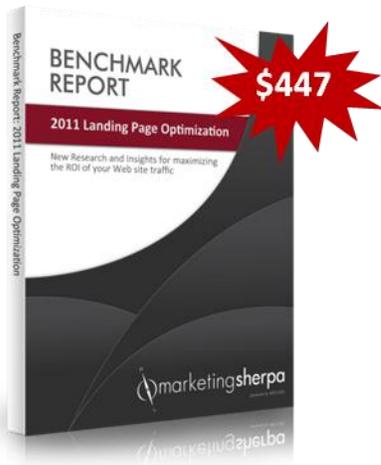
**Control:**



**Option 1:**



**Option 2:**



**Option 3:**



**Austin McCraw:** The copy, this was a very interesting part of it.

Variable: Copy

Control (long copy):

Landing page optimization (LPO) is becoming more affordable as it grows in sophistication. In fact the ROI for successful LPO programs is impressive.

According to MarketingSherpa, to be successful LPO must have an objective. That can be as simple as generating a lead or closing a sale. But instead of looking only at the business aspects, you should also frame your objectives in terms of human behavior. In other words, LPO is optimization of websites for specific human behaviors.

The chapter *Key Components of a Successful Landing Page Optimization Strategy* is about what your website visitors want rather than how to drive them to your website. In it you will find out the different methodologies marketers employ to:

- Determine visitor motivation
- Generate relevant experiences
- Create a model for quantitative evaluation

**Based on data from 2,673 marketers, this chapter will give you quantifiable information on:**

- The use and effectiveness of dedicated landing pages vs. your default website
- The value of competitive intelligence
- The kinds of metrics collected, and the effectiveness of visitor data in optimizing for relevance
- Lead quality scores and who is using it
- Marketing insights on how to create relevant and targeted pages

**Start maximizing the ROI of your website traffic today. Download your free copy of *Key Components of a Successful Landing Page Optimization Strategy* by filling out the form to your left.**

Option 1 (short copy):

Landing page optimization (LPO) is becoming more affordable as it grows in sophistication. In fact the ROI for successful LPO programs is impressive.

According to MarketingSherpa, to be successful LPO must have an objective. That can be as simple as generating a lead or closing a sale. But instead of looking only at the business aspects, you should also frame your objectives in terms of human behavior. In other words, LPO is optimization of websites for specific human behaviors.

The chapter *Key Components of a Successful Landing Page Optimization Strategy* is about what your website visitors want rather than how to drive them to your website. In it you will find out the different methodologies marketers employ to:

- Determine visitor motivation
- Generate relevant experiences
- Create a model for quantitative evaluation

**Option 2 (bullets only):**

Based on data from 2,673 marketers, this chapter will give you quantifiable information on:

- The use and effectiveness of dedicated landing pages vs. your default website
- The value of competitive intelligence
- The kinds of metrics collected, and the effectiveness of visitor data in optimizing for relevance
- Lead quality scores and who is using it
- Marketing insights on how to create relevant and targeted pages

**Option 3 (chapter excerpt):**

Getting a human being to behave in a certain way requires pulling the right psychological levers and setting off appropriate triggers.

Within the context of optimizing websites, this means communicating the offer in a way that matches the visitor's preferences and motivations both for the offer itself and for the process by which the visitor can accept the offer (make a purchase, fill out a lead form, etc.). "Relevance" is a term, with which marketers are familiar, but typically it refers to the match between Web page subject matter and the visitor's expectations.

In a broader sense, LPO is about creating a match with visitor preferences and motivation, not only in the subject matter, but also in the value being offered, the way that value is communicated (messaging), and the process by which the interaction between the website and the visitor unfolds (experience) to increase the likelihood of conversion.

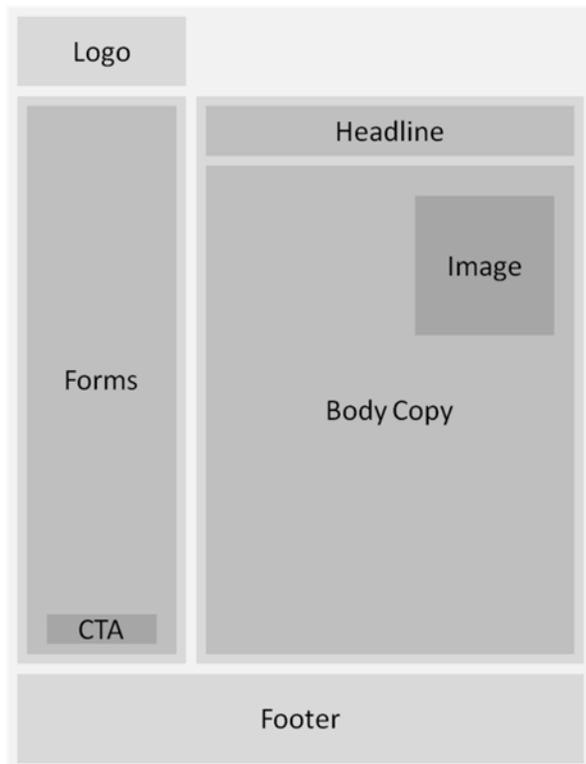
While this makes sense in theory, the reality is that creating this match demands that marketers obtain, analyze and apply deep insights into the website visitor preferences. Understanding the visitor to create relevant and valuable experiences becomes the foundation of an LPO strategy...

*- Excerpt from the Landing Page Optimization Benchmark Report*

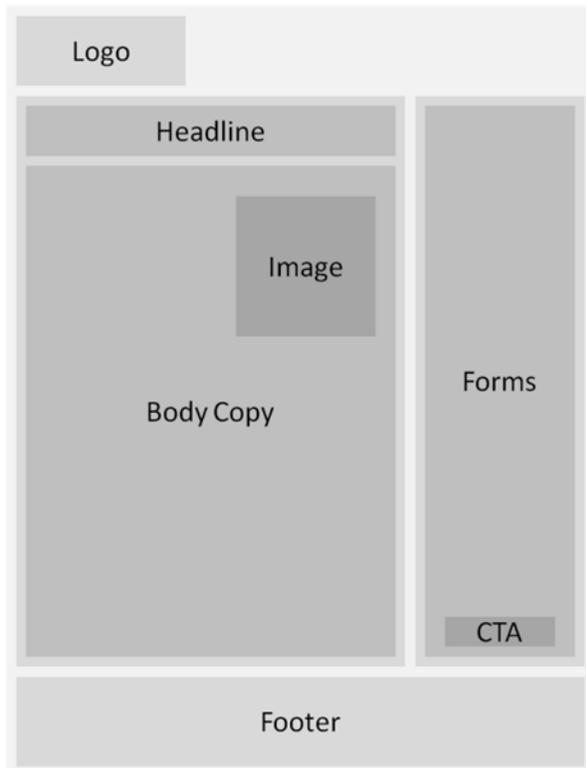
**Austin McCraw:** We gave them four options of copy as well. The control was a long copy. We gave them the option that was short copy. There were two other options. One was just bullets only. Then, we kind of had this radical option where it was just an excerpt of the chapter and it would say, "Read more." They chose ... actually went with short copy for that.

**Variable:** Layout

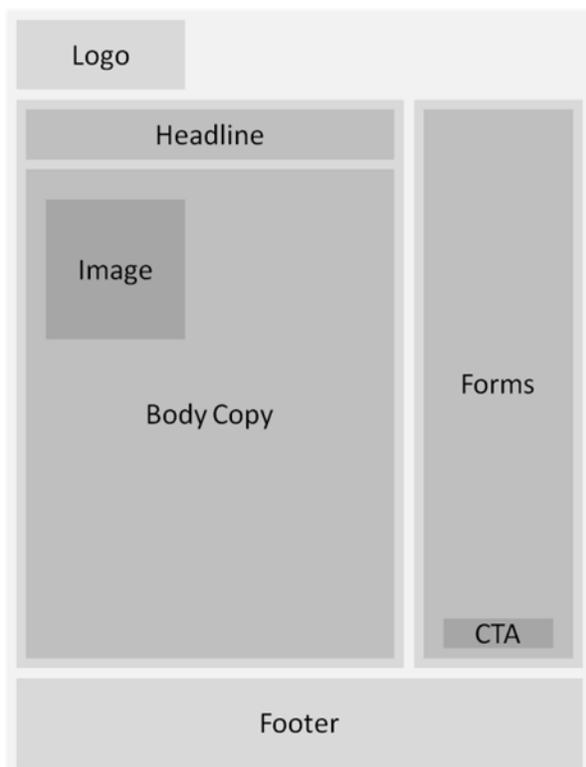
**Control:**

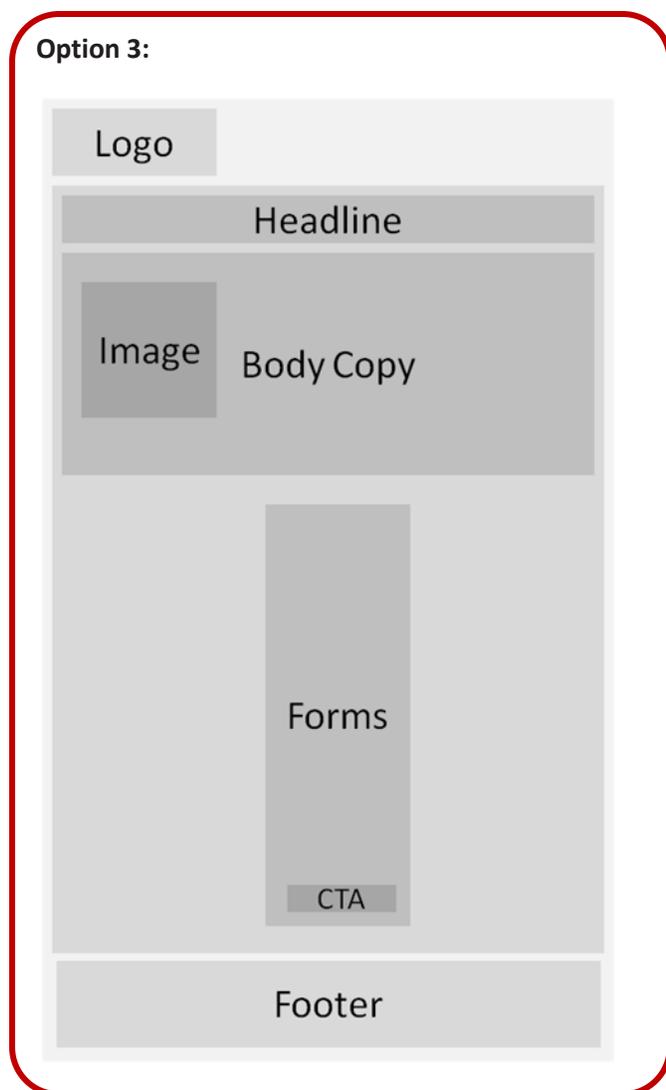


**Option 1:**



**Option 2:**





**Austin McCraw:** And, the final variable that we let them test here was the layout of the page. You can see two layouts, control, Option #1, and there were two more that they were able to choose from, 2 and 3, and they chose Option #3.

## Experiment: Treatment

Top



Get a FREE 40-page Chapter Containing the Latest Landing Page Optimization Research from MarketingSherpa

Discover the Key Components of a Successful Landing Page Optimization Strategy



Landing page optimization (LPO) is becoming more affordable as it grows in sophistication. In fact the ROI for successful LPO programs is impressive.

According to MarketingSherpa, to be successful LPO must have an objective. That can be as simple as generating a lead or closing a sale. But instead of looking only at the business aspects, you should also frame your objectives in terms of human behavior. In other words, LPO is optimization of websites for specific human behaviors.

The chapter *Key Components of a Successful Landing Page Optimization Strategy* is about what your website visitors want rather than how to drive them to your website. In it you will find out the different methodologies marketers employ to:

- Determine visitor motivation
- Generate relevant experiences
- Create a model for quantitative evaluation

Get Your Free Chapter Now!

---

First Name \*

Last Name \*

Email ([privacy policy](#)) \*

**Austin McCraw:** So, combining that all together, this is the page that they produced. You can see the headline. You can see the copy. You can see the images.

**Bottom**

First Name \*

Last Name \*

Email (privacy policy) \*

Phone \*

Company \*

Company Website \*

Role at Company \*

- Please Select -

Number of Employees \*

- Please Select -

Does Your Business Primarily Sell to Other Businesses (B2B) or Consumers (B2C)? \*

- Please Select -

Does Your Business Provide PR, Web Design, SEO, or other Marketing Products or Services? \*

- Please Select -

Biggest Marketing Challenge

Sign me up for MarketingSherpa's Inbound Marketing eNewsletter.

**Get Your Free Chapter Now**

---

[Tweet This](#) [Share on Facebook](#) [Share on LinkedIn](#)

**Austin McCraw:** If you go down to the bottom of the page, you can see the call-to-action. So, they just created it side by side.

## Experiment: Side-by-side

Control

Treatment

**Dr. Flint McGlaughlin:** All right. So, taking a look now, you should be able to see the control, and you should be able to see the treatment. And, I'd like to just take a moment and I'd like to exclude you if you were at this event so that we could get a fairly accurate sampling from our audience. But, I'd like to invite you to use the Q&A feature to vote for the control or the treatment. I just want to warn you, there is no trick here. The statistical difference will be small, and we'd like you to vote right now to tell us which one: A or ... control or treatment. Just type it in. Sally says it's the control. Treatment says Brian. Treatment says another. Control. Treatment. Control. Control. Treatment. Control. The audience is quite divided on this. It looks to me like about a 50:50 set of votes coming in. Treatment. Control. Treatment. Treatment. Control. Control.

**Austin McCraw:** Now, you know how hard it is to get marketers to agree on a design.

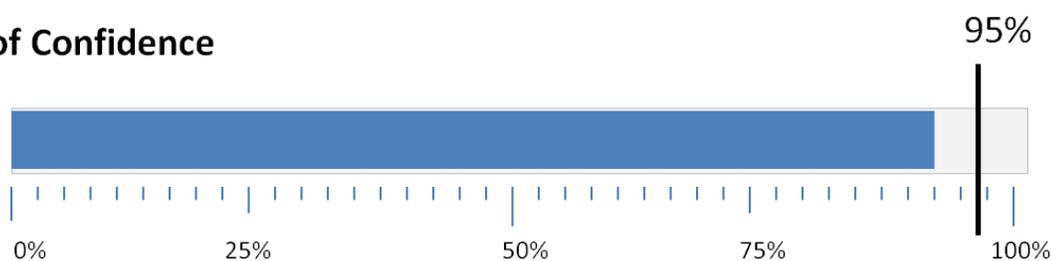
**Dr. Flint McGlaughlin:** Yeah. It's ... one of the things that we've done, while you're votes are coming in, that I think you'd find interesting is we've often shown the marketers these kinds of samples before, and used polling features, and have them tell us which particular treatment will be the winner or which particular design. And, in almost every case ... in fact, we ran two major studies on marketer intuition, and in the first one over 70% of the marketing audience chose the wrong control or treatment. And, the second test that we ran, almost a year later, produced almost the same error ratio. It's quite

fascinating. It only points to the need to test. But, this is particularly hard. So, if you voted control or treatment, we understand. Now, I will tell you this. Overall, I think the treatment has a better look and feel than the control, but with motivation levels this high, it hardly matters. In fact, what would be interesting would be to add form fields and the test I'd want to run would be to see what I could get away with, without impacting conversion. It might be that when you have a level this high of motivation, the best thing to do is to actually increase your friction and try to get more accomplished in terms of your customer theory. That is, you know, learning about your customer so you can make more intelligent decisions in your marketing. Are you ready yet to see the results? We want to show them to you because they set up much of what you need to learn today. So, here is the data set.

### Experiment: Results

Versions	CR	Rel. diff	Stat. Conf
Control	47.91%	-	-
<b>Treatment</b>	<b>48.24%</b>	<b>0.7%</b>	 <b>90%</b>

### Level of Confidence



**Austin McCraw:** Well, the treatment improved conversion by a whopping 0.7%. And, again, like Flint was saying, the motivation levels are pretty high. A 1% increase on this would be great! We've seen 1% conversion rates, or increases, produce millions of dollars in the past.

**Dr. Flint McLaughlin:** Yeah. Yes.

**Austin McCraw:** I will say this. And, the reason why you see that Level of Confidence bar in the bottom is because when we started this test, we decided that we wanted a 95% statistical level of confidence.

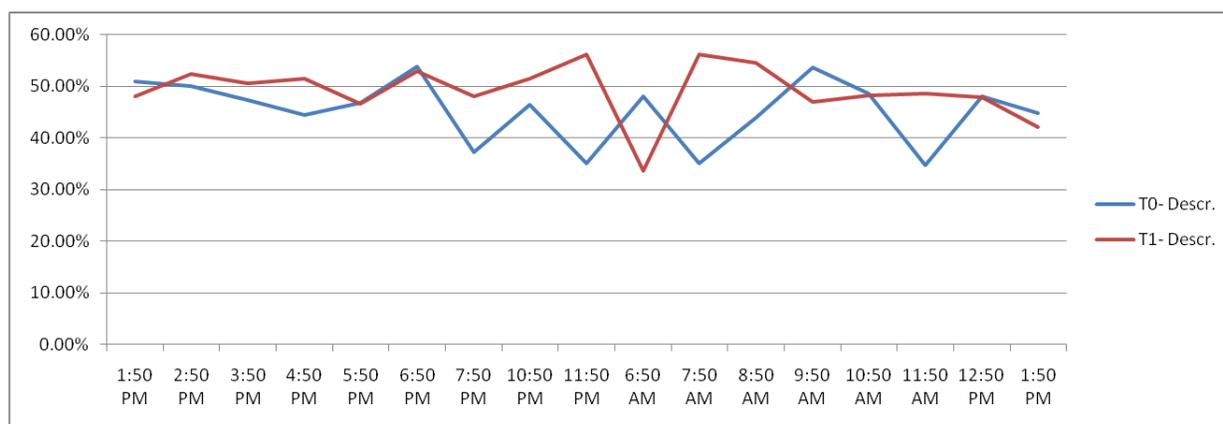
**Dr. Flint McLaughlin:** Yes.

**Austin McCraw:** And, at this point, at the end of the test, all we had was a 90%.

**Dr. Flint McGlaughlin:** Now, if you're tuned in and this is your first event, you're liable to be saying to yourself this is, you know, not that dramatic a difference. If you've been to any MECLABS events, MarketingExperiments events before, this is probably the smallest lift we have ever demonstrated. But, the point is not the lift. We have only told half the story. So, listen carefully. In fact, watch what I'm going to show you next.

### Experiment: Data

#### Conversion Rate (page views-to-downloads)



**Dr. Flint McGlaughlin:** This data represents the two designs: the control and the treatment. The control is the blue line. The treatment is the red line. And, I want you to notice at the bottom is the actual day. This was a very, very short test and ... because it had to be. We were announcing the results of the event, and so we needed a test that ... and get enough traffic that we could actually reach a level of confidence in a short period of time. And, hour by hour, you should start to see 150, 250, 350. This is from the beginning, all the way through. And, watch what's happening. Watch while one is winning, and then the other is winning. And, this is going out throughout the day. Now, if you are experienced with data, when you see these two lines represented in graph form, something should be bothering you. In fact, I invite you to look right now and see what you can determine that might indicate there is problem somewhere in the design of this test. While you're thinking about that, I'm going to go back to Austin, and I'm going to let Austin just kind point out what we see in the underlying data.

**Austin McCraw:** Yeah. Let me just also say something kind of personal here. I'm normally working with content here at MarketingExperiments. I rarely get to kind of get my hands in, dirty, running the tests. This is like one of the first times I got to do it. For me, this was kind of like a roller coaster ride. I mean, it was like when we first launched the test, the treatment was winning. It was like a 3% increase in conversion. It was like 87% confidence level. We went to bed that night. We came back the next day and the results had completely flopped. And, so the reason I say that is because that ... when you have something like that happen, and when you have your results flopping like that, that's generally ... and

what I've learned was it's often an indicator that there's a pretty decent validity threat with your test. Something is going on that is affecting your test, that's affecting the results.

**Conversion Rate (page views-to-downloads)**

**Important Usage Note:** The Test Validation tool reads sample data directly from the table below, so you should use caution if alteration of the table structure is necessary in order to unambiguously reflect the test conversion path. Please contact the Director of Sciences with any testing situation that cannot be accommodated within these constraints.

Figures used for Validation	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7	Period 8	Period 9	Period 10	Period 11	Period 12	Period 13	Period 14	Period 15	Period 16	Period 17	Period 18
<b>Validation Total</b>	1:50 PM	2:50 PM	3:50 PM	4:50 PM	5:50 PM	6:50 PM	7:50 PM	8:50 PM	9:50 AM	10:50 AM	11:50 AM	12:50 PM	1:50 PM	2:50 PM	3:50 PM	4:50 PM	5:50 PM	6:50 PM
	8%	4%	13%	8%	6%	-11%	5%	28%	3%	-5%	44%	-5%	-3%	-13%	13%	-1%	-12%	-43%
	4784	1167	920	461	308	235	162	140	319	77	435	77	91	95	103	75	52	67
	4784	1167	920	461	308	235	162	140	319	77	435	77	91	95	103	75	52	67
	2292	595	460	218	137	110	87	52	148	27	209	27	40	51	50	26	25	30
	2292	595	460	218	137	110	87	52	148	27	209	27	40	51	50	26	25	30
	47.91%	50.99%	50.00%	47.29%	44.48%	46.81%	53.70%	37.14%	46.39%	35.06%	48.05%	35.06%	43.96%	53.68%	48.54%	34.67%	48.08%	44.78%
		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	5178	1209	1042	497	327	210	170	179	328	73	625	73	88	83	116	74	46	38
	5178	1209	1042	497	327	210	170	179	328	73	625	73	88	83	116	74	46	38
	2498	581	546	251	168	98	90	86	169	41	210	41	48	39	56	36	22	16
	2498	581	546	251	168	98	90	86	169	41	210	41	48	39	56	36	22	16
	48.24%	48.06%	52.40%	50.50%	51.38%	46.67%	52.94%	48.04%	51.52%	56.16%	33.60%	56.16%	54.55%	46.99%	48.28%	48.65%	47.83%	42.11%
		-2.93%	2.40%	3.21%	6.90%	-0.14%	-0.76%	10.90%	5.13%	21.10%	-14.45%	21.10%	10.59%	-6.70%	-0.27%	13.98%	-0.25%	-2.67%

**Dr. Flint McLaughlin:** So, let's look at the data. If you were to get past those graphs and dig down deep, you would see something like this.

**Austin McCraw:** Yeah. And, what we learned here, what we noticed as we kind of tried to pinpoint those moments where the conversion rate dropped for the treatment particularly, what we saw in the data is that ... you know, we were supposed to be splitting the traffic 50:50, but in those periods of time, what happened was for some reason all of the sudden the treatment was getting 28%, 44% more traffic than the control. And, then at the end you can see the control ended up getting, you know, 43% more traffic than the treatment.

**Dr. Flint McLaughlin:** So, because this is one of our more technical clinics, what you need to be clear, that you understand in this chart. The line, the white line with the red circles on it is actually representing the traffic coming to the various treatments.

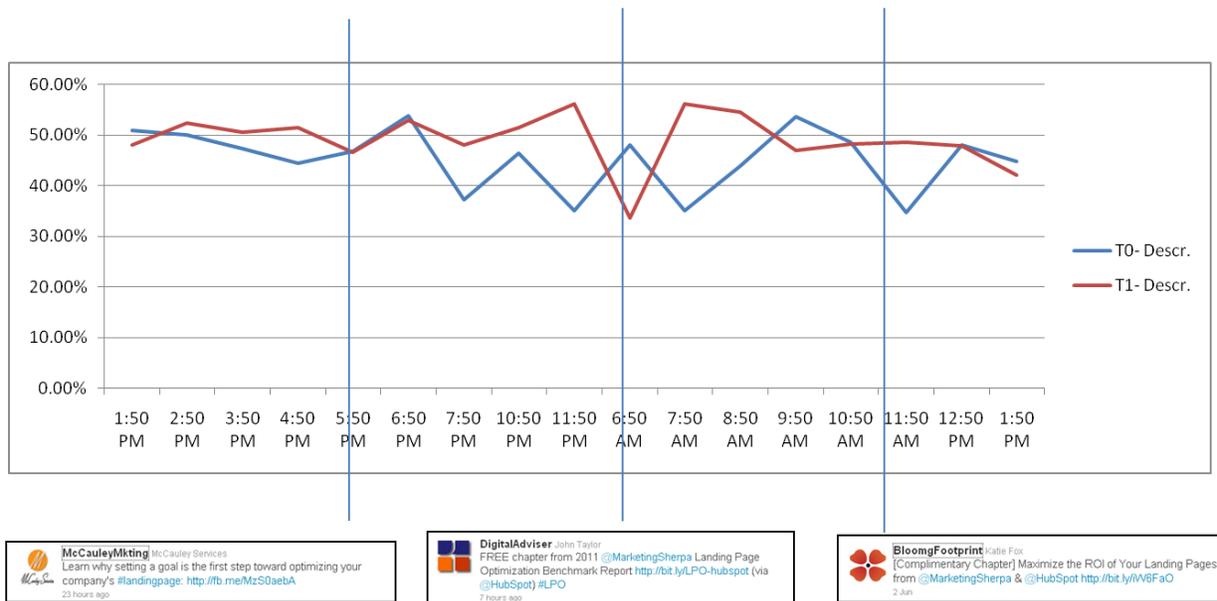
**Austin McCraw:** The split of the traffic. Yes.

**Dr. Flint McLaughlin:** It's to split itself. It's the differential. And, you see the big swings in traffic. Now, that led to a problem. That led to us realizing that many people were eager to see the results of this test, waiting for us to announce it, while we simply could not and would not make any pronouncement that would lead to an announcement unless we understood what was going on in the data set. So, what was happening behind the scenes ... and I'm going to take you to a final slide and let Austin tell you so we can move on, because we have a lot to teach you today.

**Austin McCraw:** Yeah.

**Dr. Flint McLaughlin:** There is learning in this as well.

### Conversion Rate (page views-to-downloads)



**Austin McCraw:** Well, just trying to figure this out, what's going on, we looked at possibly the splitter, having an issue with it. It did not. It was running perfectly. What really had happened, and this is partly even my fault here as kind of a newbie here at testing, is that we had not controlled the URLs for the control and the treatment. We were directing the traffic to a splitter page, which would then split 50% of the traffic this way, to the control URL, 50% of the traffic to the treatment URL. What happened from there is that people, if they landed on a treatment, they liked the offer, they wanted to share it with their friends, they would then tweet that URL. And, so all of the sudden you get this flood of new people coming in through tweets, through blog posts, from outside of our test. They weren't going through the splitter. And, ironically ... and the way that affects the traffic is, and what we discovered, which is interesting learning here, is that the motivation levels for kind of the second-hand traffic was a lot lower than the motivation for the first-hand traffic. And so, the conversion rate would drop whenever the treatment received more traffic.

Versions	CR	Rel. diff	Stat. Conf
Control	47.91%	-	-
Treatment	48.24%	0.7%	 90%



**What you need to understand:** Many marketers consider a 90% level of confidence conclusive. However, had we considered this test conclusive and pushed the winner live, it would have been a poor decision.

**Austin McCraw:** So, even though we had a 90% confidence level here, even though ... and there are some tools that will ... if you get a 90% confidence level, it will say you're good to go. Launch the treatment; it won. Even though we've had that here, looking at the data more specifically, we really pinpoint a pretty significant validity threat that we had to really throw out the test results. You know, we couldn't make any conclusions.

**Dr. Flint McGlaughlin:** They were not conclusive. Absolutely! And, that's part of being in the position we're here in here. At MECLABS, and MarketingExperiments in particular, we say every marketer should have at least two key virtues. The first one is a brutal honesty. You've got to be able to look at the numbers and even if they aren't what you want them to be, you have to admit what they're saying. That is not easy. I've seen many marketing groups that struggle with that level of honesty. That means nobody gets to be the expert guru all of the time, because anybody that's running a good test design is going to have their basic assumptions challenged and they're not always going to be right. In fact, if every test produces a lift, you're not testing right. You're not digging down deep enough into customer motivations, you're not understanding enough to improve your customer theory, and so you are perhaps looking smart to your colleagues but you have left way too much money on the table. The greater your customer theory, the greater your conversion rate. It's not just the test. The test is not to learn. So, brutal honesty is the first trait. And, the second one is a kind of basic humility. You've got to be able to say, "We screwed up," and you've got to be able to say it a lot. And, you've got to be able to say, "I thought X was going to happen," when Y happens, and say, "I don't even understand it." In this case, you have a whole team of experts, you have scientists, you have analysts, but we still have an inconclusive test. And, that happens, and you'd better be prepared for it, or you're liable to take your company in the wrong direction.

---

## Landing Page Optimization Workshop

---



**Get Certified in Landing Page Optimization Today!**

marketing experiments  
2011

LANDING PAGE OPTIMIZATION CERTIFIED

- 1) Earn Professional Certification in Landing Page Optimization
- 2) Receive a Free 2011 LPO Marketing Benchmark
- 3) Get Live Personalized Training from LPO Experts

Coming to DC - Nov. 8

[Learn More About the Workshop >>](#)

## What we discovered

---



### **Key Principles**

1. Just because a test *looks* conclusive doesn't mean it is **conclusive**.
2. There are at least 3 validity threats beyond sample size that you need to consider when testing:
  - History effect
  - Instrumentation effect
  - Selection effect

**Dr. Flint McGlaughlin:** In the meantime, there are three things that you need to learn in order to protect yourself, and I'm coming towards it. Here are the principles. Here is a summary of what I've said thus

far and what Austin has been demonstrating. Just because this looks conclusive doesn't mean it is conclusive. And, there are these three ... now, there are other threats. In fact, if you take one of our certification programs, we'll teach you the whole set of effects, but these are the three most common effects. In fact, if you can learn these three on this clinic, you can perhaps save yourself a lot of mistakes and potentially a lot of money. And, it's possible to learn enough in the next few minutes we have together to actually make this happen. So, let's move swiftly to teach the first, history effect.

### **Validity Threat #1: History effect**

---

Before I start to break it down, I'm going to ask the audience a question, so get your Q&A ready. And, I want you to tell me what you believe history effect is. If you're certified, just put a C in there so we know it's coming from one of our students. If you're not, don't put the C in there. And, let me see what the audience thinks history effect is, as you're planning a test. Meanwhile, I see your questions and our tech team is standing by. And, by the way, while your answers are coming in, I was working with one of our videographers this morning. We've built a pretty cool video about what it takes to produce one of these clinics, and it's going to be released soon on our site. Luke is with us today, and he is controlling sound as this takes place. But, it goes behind the scenes and shows you all of the people working and all of the ... literally the hundreds of hours that are invested in our clinics, and shows you the studio. And, of course, if you see me now, I'm, as always, in a formal suit — not— wearing dress shoes — not. You'll get to see what it looks like when we're actually all sitting around here. Most of us are in flops in Florida. And, I will wear my suit at the Summit, but I don't wear it here very often, and so watch for that video. We'll release it soon. And, we should probably announce to the audience when we're releasing it, instead of leaving it just on the site, which I think is the plan, but we can talk about that.

In the meantime, here are your thoughts. History, how fresh are your prospects? What have you previously sent them? How recent? Good guess, Eve! That's not the technical definition, but I'm glad. Means to look at browsing history. Have a similar test performed in the past. Marco says “dunno,” D-U-N-N-O. And, so expert visitors versus regular visitors. External factors. Oh...oh! Geoffrey, you're getting close. Let's keep going, all the way down. Excellent! I'm looking at your responses. Good! All right. The audience is helping me as a teacher, because I'm not in a room. And, I am a teacher at heart. All I want to do really is do research, conduct content and teach. Unfortunately, I have to direct an organization that gives us the capacity to do that, but it only enables us. And, when I look at what you're telling me right now, in the audience, I see the need for what we're talking about. And, it's what I expect. And, if you don't know the answer, don't feel bad because you can get your MBA right now at Duke or Wharton in marketing and not know this. And, I've talked to Ph.D.'s in testing who don't know these things. But, what I'm going to show you right now, on the history effect, isn't something developed in our lab. This is classic testing theory with applications from our lab. And, so let's work on it.

### **History Effect: Definition**

---

**History Effect:** The effect on a dependent variable by an extraneous variable associated with the passing of time.

**Plain English Definition:** Something happens in the outside world that causes flawed data in the test.

**Dr. Flint McGlaughlin:** First of all is history effect. Here is the official definition of history effect. It's the effect on a dependent variable by an extraneous variable associated with the passing of time. That is the definition that is rich with meaning and also meaningless for those who haven't taken the time to work it out and parse it a word at a time. I think that our writer, Paul ... where is Paul? Paul is in the room here somewhere. I see him at the back. Are you monitoring Twitter, Paul, or what are you working on? What's that?

**Paul Cheney:** The Q&A.

**Dr. Flint McGlaughlin:** Q&A. All right, Paul is monitoring Q&A, but he is the writer who helped produce this particular clinic. And Paul is a really good copywriter, and his definition, is "something happens in the outside world that causes flawed data in the test." Now, I don't think that'll pass the exam, but I think that will certainly helps our audience understand in plain English what's going on here. And, we're teaching you both because we really want you to have a level of expertise and recognize this, but you can focus on the second definition just to get to the pragmatic side of how do I make this happen, how that works. So, you get the idea that something from the outside is happening. It's happening in time. And, because of what it's doing, it's skewing your results, or potentially skewing them.

With that in mind, let's look at a precise example. I'm going to move faster now. My voice is going to pick up speed. That's deliberate, so bear with me. If I go too fast, I'll slow down. But, we've set it up now, and so now I want to deliver as much as I can.

### History Effect: Example

---



**Experiment ID:** *Protected*

**Location:** MarketingExperiments Research Library

#### Research Notes:

**Background:** Online sex offender registry service for parents concerned about their surrounding areas

**Goal:** To increase the click-through rate of a PPC advertisement

**Primary research question:** Which ad headline will produce the most clickthrough?

## Test Design: A/B/C/D split test focusing on the headlines of a PPC advertisement

**Dr. Flint McLaughlin:** This is an experiment from our test library. It is an old experiment, and I really like it, and I remember it. Online sex offender registry service, this is back when those first started coming out, and we had one that we were working with. And, the goal was to increase the clickthrough rate of a page search advertisement. This is a service that allows you to see the names and the criminal record associated with anyone in your neighborhood that might be a sexual predator. And, all you have to do is put your zip code in and they come ... you know, there is a list of the records. And, they update you when sexual predators move into your neighborhood, and so that's what the service was. We were looking for a headline that would produce more clickthrough.

We prepared a headline test using Google AdWords as the split-testing platform. The headlines were chosen by the participants of the certification course from a pool which they created. The test was conducted for seven days and received 55,000 impressions.

The screenshot shows the MSNBC website interface. At the top, there is a search bar and navigation links for 'MSNBC Home', 'Dateline NBC', and 'To Catch a Predator'. The main content area features a large image of two men in a bar setting. Below the image is the headline 'Admitted child abuser caught in Dateline sting' and a sub-headline 'To Catch a Predator': Cameras follow the potential predators from chatting about sex online, to shuffling through the criminal justice system. Plus, the most frightening man caught in this investigation. Chris Hansen reports. • FULL STORY'. To the right of the main article is a video player titled 'The next investigation' with a 'Launch' button. Below the video player is a section titled 'WEDNESDAYS ON NBC' with text about an upcoming investigation in Florida. At the bottom of the page, there is a small 'advertisement' placeholder.

### [Child Predator Registry](#)

Identify sex offenders living in your area. Protect your kids today.  
[www.XXXXXXXXXXXXXXXXXX.com](http://www.XXXXXXXXXXXXXXXXXX.com)

### [Predators in Your Area](#)

Identify sex offenders living in your area. Protect your kids today.  
[www.XXXXXXXXXXXXXXXXXX.com](http://www.XXXXXXXXXXXXXXXXXX.com)

**Dr. Flint McLaughlin:** So, we had four ads. Please look at them. "Child Predator Registry: Is your child safe? Predators in Your Area." And, "Find

Child Predators." Now, you may analyze these page search ads and try to determine which one is best. In fact, take a look. Lock down in your mind the one you think will be best. You don't have to vote, but you can. But, just take a look and kind of get a sense you think will produce. We ran a split testing

platform. The test was conducted for seven days, and we had 55,000 potential actions to measure. What does that tell us? Well, look.

During the test, Dateline aired a special called “*To Catch a Predator*,” which was viewed by approximately 10 million individuals

Throughout this program, sex offenders are referred to as “predators”

**Dr. Flint McLaughlin:** Here’s the problem. During the test, Dateline aired a special called “To Catch a Predator.” It was viewed by 10 million people. The words predator became the key term associated with sex offender. Now, let's go backward. You see is your child safe. You see find child predator, predators in your area, and child predator registry. And then, look in the copy. Identify sex offenders, identify sex offenders. All the same except for the headline, but we have three of these headlines with the word predator in them. What was the result?

Headline	Impressions	Clicks	CTR
<b>Predators</b> in Your Area	21,096	1,423	<b>6.74%</b>
Child <b>Predator</b> Registry	14,712	652	<b>4.43%</b>
Find Child <b>Predators</b>	18,459	817	<b>4.42%</b>
Is Your Child Safe?	15,128	437	2.89%

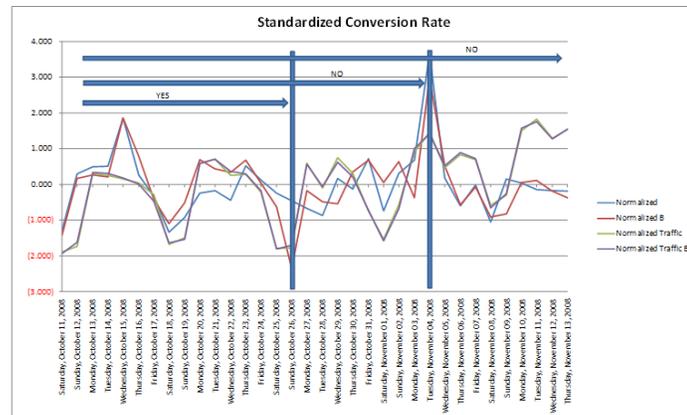


**What you need to understand:** In the two days following the Dateline special, there was a considerable spike in overall click-through, but a relative difference between those ads with "predator" in the headline and those ads without "predator" of up to 133%. **So, in effect, an extraneous variable (the Dateline special) associated with the passage of time jeopardized the validity of our experiment.**

**Dr. Flint McLaughlin:** Well, in the two days following the Dateline special, there was a spike in clickthrough, but a relative difference between those ads with “predator” in the headline and those ads without “predator” of 133%. So, in effect ... now here it is in bold, that same technical definition but it at least it’s in context. So in effect, an extraneous variable, (the Dateline special) associated with the

passage of time (that's the test cycle, i.e., time that it aired) jeopardized the validity of our experiment. That's the data you see up on your screen. This is an example of history effect. And, if you've tuned into this right now, you are probably learning history effect for the first time and not recognizing that a lot of the activities going on around you in the media, on YouTube, and throughout the Internet can have an impact on the validity of you tests. Some of you who are in publicly traded companies whose names are constantly in the press, in the paper, and not realizing that even the press itself on the outside can have a significant impact on what's going on within your testing.

Graphed results of a 4-week email test with an ecommerce retailer:



## History Effect: Precautions

- Make sure everyone in the company knows you're testing
- If possible, track day-to-day data
- If possible, always run a/b split tests
- Use media tracking tools (Google alerts, etc.) if you plan on testing around search terms impacted by the media
- Monitor for test anomalies

**Dr. Flint McLaughlin:** Now, you probably want to know what you can do to prevent this, and I don't have 8 hours to teach. And, so what we've done is extract practical, simple things that you can do to protect your tests against history effect. First of all, you need to make sure everyone in your company knows that you're testing. At least if it's a small organization make sure everyone knows. If it's a large organization, make sure the *right* people know. If you don't, you can have activities taking place at the

same time that are hurting you. Number two, if possible, don't just wait and look at your test at the end, but track the data day-by-day so you can determine any anomalies in the patterns. Number 3, if possible always run a/b split tests. Now let me back that up, that doesn't mean don't do multi-factorial tests. Actually, a/b split, just to get technical with you, is a single factorial test. And, what you call multi-variable or multi-variance is called a multi-factorial test in testing vernacular. And we're not saying you can't run any multi-factorial; what we're saying is don't run a sequential test. Now, you may be asking what are sequential tests. A lot of people on the Internet now put the page, run it for a week, take it down, put up page 2 or the treatment, run it for a week and then compare the numbers. Avoid them if you can! This brings me to another point, use media tracking tools if you plan on testing around search terms impacted by the media. We had to set up alerts that would allow us to track the word predator and sexual offender. And, we found that when we were running experiments regionally that there might be a big story somewhere in Minnesota that impacts all of the paid search results within Minneapolis, and so we had to be careful as we were conducting our test. But, alerts can help you be aware of these threats as they come in.



And, number four, monitor for test anomalies. Look at this data set in front of you. If you were to watch it very carefully, it's a four-week email test with an e-commerce retailer, and in the first week, you can see the pattern, but something goes wrong in the second week, and there is a dramatic difference. Watch that. As that begins to happen, it tells you there's something going on that's impacting your results. This is a good visual. Later, in fact if you were to come to one of our certification classes, we'll take these graphs and show you four of them side by side, and let you start picking up the one with the potential problem so that you can visually see it, as well as go down to the data set and discern it. So, then, we're going to move on.

### Audience Question

---

**Seasonality. Does a valid test in the "off-season" translate to our busy season? -Greg**

**Dr. Flint McGlaughlin:** But, someone has asked, "Seasonality, does a valid test in the off-season translate to our busy season?" And, I've got an expert in that area who will answer us, and I'm going to take you right after this to the second validity threat. Go ahead. This is Phillip.

**Phillip Porter:** That's a good question, Greg! The short answer is that sometimes. Last year, we had run a test with a partner whose business is very seasonal, and the test started right before Christmas and then ran for several weeks after Christmas. Right before Christmas, the visitors were very motivated to make a purchase and the treatment wasn't very different from the control. After Christmas, the visitors were less motivated. Flint talked about how important motivation is in heuristics. And, the treatment

performed much better than the control. Generally, findings from tests which address motivation will translate less from the off-season to the busy season, and finally, related to things like clarity, the value proposition, friction, anxiety, will translate better. But, keep in mind that nothing can replace testing during the various seasons. And, so if you've identified a seasonal pattern in your business, you need to replicate any off-season testing during your busy season to make sure that those results are valid for the different time periods. Sometimes, the results will translate to your busy season, but you don't know for sure until you run a test during your busy season.

**Dr. Flint McGlaughlin:** I would pre-test before the season, but let's take holidays. Christmas is coming up. You do realize that paid search traffic has higher motivation in the holiday season, and I have seen people, and this is where they really get in trouble, I've seen them take the test results during the holiday season and make decisions for January and February based on them, not realizing that the intense motivation at that time of the year is skewing their results and will impact them so that when people have less motivation, suddenly anxiety, friction and problems with the value proposition become much more impactful on the conversion. Phillip, I have seen him, by the way, take a whole, huge, I'll call it mess when it comes to numbers, and come back and ask these penetrating questions about seasonality that I...you know, you wonder how he figured out there was seasonality, but it was all there, hiding in the data set and we didn't pay attention to that. Good question, Greg. We're moving on!

## Validity Threat #2: Instrumentation effect

---

**Dr. Flint McGlaughlin:** Let's look at Point 2, instrumentation effect. How many of you are instantly familiar with instrumentation effect? More people can guess this one than historic, but I'm going to take you straight through a classical definition.

**Instrumentation Effect:** The effect on the dependent variable, caused by a variable external to an experiment, which is associated with a change in the measurement instrument.

**Plain English Definition:** Something happens with the testing tools (or instruments) that causes flawed data in the test.

**Dr. Flint McGlaughlin:** This is the effect on the dependent variable caused by a variable external to an experiment which is associated with a change in the measurement instrument. Have you got that? Thank you for joining us today. We hope that you've gotten all that you ... oh, they're laughing at me. Clearly, that's one of those beautiful academic statements that requires a lot of parsing again, and so we turn over to our interpreter/copywriter, Paul, who says it's something that happens with the testing tool that causes flawed data in the test. Thank you, Paul! I'm going to start calling him Dr. Paul, now that he has ... so, Dr. Paul tells us that everything above can be understood with the simple sentence down below. So, learn that, and let's do the same thing we did on the last point. Let's look at an example, and then let's learn how to prevent it. Before I go there, will the audience give me some feedback with the Q&A? Tell me if I'm going at the right pace for you. Am I too fast? Is this just right? Are you learning? Are you liking this? I need to optimize my presentation live, based on your feedback. And, since you're

not in the same classroom with me, I need to see your response. So, Paul or Luke, let me see how the audience is responding and just quickly tell me with the Q&A feature, how this is going for you. Awesome! Awesome! Good, thanks! Keep going. Excellent! Etc. Etc. All right, good! It looks like positive feedback. Will my team continue to monitor that and look for any negative feedback? And, let's keep rollining.

### Instrumentation Effect: Example

---



**Experiment ID:** *(Protected)*

**Location:** MarketingExperiments Research Library

#### Research Notes:

**Background:** Online sex offender registry service for parents concerned about their surrounding areas

**Goal:** To increase the conversion

**Primary research question:** Which landing page will produce the most clickthrough?

**Test Design:** A/B spit test (variable cluster)

**Dr. Flint McGlaughlin:** All right, so I want to take you then to another example, and it is a completely different test but it is with the same group. It's the sex offender registry for, again, that same service. And, the goal here is to increase conversion and we have two landing pages, and we're looking for the landing page that will produce the most clickthrough.

**Control Values**

**Treatment Values**

- In this test, we were comparing five variables with two differing values each (highlighted above).

**Dr. Flint McLaughlin:** So, we have the control and the treatment. And, you'll notice there are boxes. Those boxes are not part, at least the red, and blue, and green outlines, are not ... and it looks like purple. Red at the top. So, see there where it says the values? And, then the blue, and then there is kind of a green, and a purple, and a red. Those are the areas, the variables. We're doing something in this test unique. It's a single factorial design, which means it looks like an A/B split test, but it's deploying variable clusters, which means in some ways it feels like a multi-variable test. One of my favorite ways to test is using single factorial with multi-variable clusters and then isolating within those clusters in the subsequent test. Remember, the goal is to learn the most. This is one of the ways to learn the most, fast, about you customer. Keeping that in mind, we discovered that in the testing, as this began ... and I want to show you this because here is the actual, you know, the test itself and we're excited about what's going on, and then we have a discovery that starts to skew our understanding. This is really hard for me to explain, and I don't have the oratory skills of Austin, so I'll do my best. But, Austin, let me come in behind you. Tell us what happened.

### Experiment Notes:

We discovered that in the testing software, a "fail safe" feature was enabled specifically that, if for any reason the treatment page was not running correctly, the page would default back to the control page.

- This was enabled by loading hidden Control values of experimental variables whenever the Treatment was loaded.

- This caused the Treatment pages to have substantially longer load times than the Control.

**Austin McCraw:** Well, essentially ... and this is a classic ... this is like you said earlier; this is an older research partner we were working with. It's an older test. But, what we discovered is that there was a fail-safe issue for this test. The way that the tool worked is that if, for some reason, the treatment did not work, did not run, did not load, that it had the control kind of in the background, ready to put that up for the visitor. So, there was always this backup option for the control to show if the treatment did not. And, what that meant was that it was storing, you know, five other elements at the same time. The control didn't have that issue because it was just the control. But, every time the treatment was shown there were five other elements that would have to load at the same time. And, so what happened, and again this was an older test, is that it affected load times significantly. And, I think the next slide shows the differential on the load times.

Page Load Time Reference Chart (in seconds)							
Page Size(kb)	Connection Rate						
	14.4	28.8	33.6	56	128 (ISDN)	1440 (T1)	
Control Page	50	35.69	17.85	15.30	9.18	4.02	0.36
	75	53.54	26.77	22.95	13.77	6.02	0.54
	<b>84.9</b>	<b>60.61</b>	<b>30.30</b>	<b>25.98</b>	<b>15.59</b>	<b>6.82</b>	<b>0.61</b>
Treatment Page	100	71.39	35.69	30.60	18.36	8.03	0.71
	125	89.24	44.62	38.24	22.95	10.04	0.89
	<b>137</b>	<b>97.80</b>	<b>48.90</b>	<b>41.92</b>	<b>25.15</b>	<b>11.00</b>	<b>0.98</b>
	150	107.08	53.54	45.89	27.54	12.05	1.07
	175	124.93	62.47	53.54	32.13	14.05	1.25
	Add. Load Time (s)	<b>37.19</b>	<b>18.60</b>	<b>15.94</b>	<b>9.56</b>	<b>4.18</b>	<b>0.37</b>

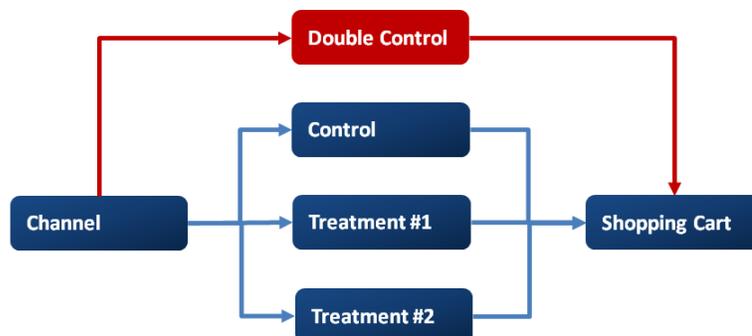


**What you need to understand:** The artificially long load times caused the “user experience” to be asymmetric among the page versions, thereby threatening test validity.

**Dr. Flint McLaughlin:** So, look carefully because you're looking right now at actual load times, and between the control and the treatment page. And, Austin's explanation is very precise and it's exactly what happened. I remember this and I remember our team trying to frantically figure out, first of all, what was going on. Because, sometimes you just know from experience that, you know, version B should probably produce a lift of X, but you're always, you know, between those two qualities of a marketer, you're ready to be show them you're wrong, but you've really got to be proven you're wrong before you adjust your whole meta-theory. And, what's going on is like, "No, this shouldn't be," and

then we saw the load time issues. And, you can see the much longer load times of the treatments by looking at these two red boxes over the control. Anything you want to add to that, Austin?

**Austin McCraw:** No. You can just see at the bottom that, I mean, depending on what kind of connection they had, they could have up to ... I mean, they could have up to like 30 seconds delay. Like, I mean on average, I think it was about a 10 second delay.



**Dr. Flint McGlaughlin:** Right.

**Austin McCraw:** But, I mean, depending on their connection, they could have a pretty significant delay.

**Dr. Flint McGlaughlin:** So, what you'd see in the testing tool is that many more people clicked away from the treatment and didn't buy, when actually they never even read the treatment, and so your data gets skewed.

### Instrumentation Effect: Precautions

---

- Set up second, backup metrics tool
- Match results to transactional data
- Test with a double control
- Monitor anomalies

**Dr. Flint McGlaughlin:** So, the question is how do I keep from having an instrumentation error? I could tell you story after story. I want to say to you that this is the most common error. I think it's going to happen more frequently than your historic effect, if you're setting up everything right and testing in tight cycles. Now, in some companies, with some products, that's not the case. But, for many of you, this is the one you should be fixing right now.

I did a major series of tests for the New York Times. We run, right now, 150 of their paths on our own servers, as we continue to conduct testing. And, I remember last year where we kept getting skewed results. And, what we found is that a leading testing tool —I could tell you the name. They're a big tool; they're an expensive tool. I won't tell you the name because I don't want to hurt the company. But, they had code implemented wrong on the New York Times' website, and it was skewing all of the results. You've got to really watch for this. So, the question becomes how do I protect myself. And, I'm going to go down a list of simple things you can do right now that will help protect you.

Number one, set up secondary backup metrics. A good example of this is Google Analytics. Now, the good people at Google, when I've mentioned this in the past, I thought they would like me telling people this. They don't because they want to be your primary. They don't want to be your secondary. But, I am telling you the same thing I told them, you're still a good tool to put on there so I can look at the reading Google Analytics gives me and compare it to the reading I got from my other tool, whether that's Omniture or some other metrics program.

The second thing is ... and you know, I think most of you know this but in case the audience isn't familiar with it, GA is free. It's easy to implement on your site and you can compare your numbers and look for differences. And by the way, there will be differences every time. You're not looking for the numbers to be identical, but you're looking for consistency in the differences. If one is over-reporting by 5% and consistently over-reports by 5%, you're okay. But if that starts changing all over the map, you've got a problem. But let's keep going. Match results to the transactional data. What does that mean? It means make sure that your 40% gain is showing up on the P&L. Now, it might not if you're a lead gen. I understand that. But, at the end, you're looking at the actual ... how many orders did I have today and how does that look, compared to what my test result tells me? So, you're looking in your accounting program and you're looking in your testing program, and you're saying, "Do they agree?"

The third one is test with a double control. And, a double control is a way to make certain that your data sets are accurate and to determine any differences in instrumentation. Shall we go into the design of that? Does anybody want to talk about that for a moment? All right, let's keep going. If you have questions about a double control, contact us and we'll try to help you. It might be something we need to ... if enough of you ask, we'll put it in the blog. If not, we'll just try to answer individually.

Number four, monitor anomalies. Those anomalies are some like we have demonstrated in the past, where you saw a difference in the data. Frankly, when Austin told you the story at the beginning, his first thought was there's something wrong with the splitter, because so often we find there's a problem with the testing tool, an instrumentation effect. It was not the problem. Technically, it wasn't the problem. There were some splitter problems I think, but they weren't the dominant problem. Austin has a point. Keep going.

**Austin McCraw:** Yeah. When we interacted with the analysts on this question, we asked them, "What validity threat did they experience the most?" It was instrumentation effect.

**Dr. Flint McGlaughlin:** Yeah. Now, that's across our team, running 1,200 studies this year, hundreds of experiments all over the world, and their most common problem is instrumentation effect. That's a very good point, Austin.

## Audience Question

---

What is the general quality of the A/B & multivariate tools on the market - do they really deliver valid results even at small sample sizes?

-Steve

**Dr. Flint McGlaughlin:** This is a question from Steve. "What is the general quality of the A/B and multi-variant tools on the market? Do they really deliver valid results, even at small sample sizes?" Oh! I didn't see that qualifier at the end until just now. We're going to go back to Dr. Phillip, Dr. Phil as we like to call him!

**Phillip Porter:** For the first part of the question, the quality of the tools on the market, it's generally good. The market works. Tools that don't work stop being used. Tools that work get used more. The second part of the question, about valid results even at small sample sizes, that's a little trickier. Valid results are a function of sample size and effect size, sample size being how much data you've collected, effect size being how big of a difference there is between the control and the treatments. As sample size goes up, the effect sizes that you need go down. Conversely, at smaller sample sizes, you need to get a larger effect size to be able to see significant results. As long as your test has more than five successes for each of the control and the treatment, the most commonly used tests would generally work pretty good. When you get smaller than that, you run into a lot of issues and there are some alternative tests that should be used in those situations. But, for the most part, our tests have more than five successes for each of the control and the treatment. So, generally speaking, most of the time the tools that are out there work well and they will give you valid results, as long as you've checked for some of these threats that we're talking about today.

**Dr. Flint McGlaughlin:** Now, I want to give you a caveat. I think Phillip's answer was precise and I have nothing to add to it, but I want to say that despite the fact that we have tools that can work, we don't have people that know how to work them very well. And, look, if this sounds like me pitching a product for MECLABS or something, MarketingExperiments, I'm not because I don't really have anything but a certification course and I promise you it doesn't represent 1% of our revenue. But, we do have a certification course in the fundamentals of online testing. We built it because we thought the market needed it so badly. We want you to...I don't have a slide for it. I'm not selling it. I'm just saying take it

because it'll help you and recognize that what's going on right now is you're in the middle of a phenomenon.

The Internet is a phenomenon. I've watched it from the beginning. One day, I'm going to do a story about how MECLABS came out with the oldest and largest lab in this field. And, when we first had a vision for what the Internet could be, you couldn't even put up a web page. There was no such thing. I was typing gopher streams in and beginning to sense the impact on broadcast. Some of you on this call don't know that you can go to our back library there and there are hundreds of hours of video and television footage, because we pioneered the testing of video on the Internet and even worked with the networks. And, some of the shows that you watch on TV even now grew out of our research in those areas. We never talk about that on these clinics. I'm only saying that to say this ... Over time, we've learned a lot and now we're excited to see people testing. I'm just afraid because I think of all of these marketers as my friends and some of you, I hardly know you personally, but I mean it's why we exist as a company, is to serve you. And, we're afraid that so many tests are being run right now that are being run without people who have been properly trained. And, I've talked to optimization personnel hired by a company as the director of optimization who don't know how to construct a robust test design and couldn't deliver a DOE. Now, if you're in optimization and you don't know what a DOE is, we already have a problem. But, if you can't build a DOE, don't know how to build a DOE, you shouldn't be in optimization. You should come somewhere and get trained, and then go back and help your company better. And, I don't mean that ... that may not be your fault at all. You may have gotten the job or put in the position. We just want to help. All right, let's keep going.

### **Validity Threat #3: Selection effect**

---

**Dr. Flint McLaughlin:** Selection effect. So, this is the third. I have 10 minutes. I will promise to put 15 to 20 minutes worth of content into the next 10 minutes, if you'll stay on to the last second. And, I don't get any points or any money if you do, but if you'll stay on to the last second, we'll pack it with learning. Because we have 10 minutes and because I'm going to teach right up to the end, I do want to say this. I don't want to lose this time at the end, so I'm going to say it to you right now. For years, we did these clinics with no sponsorship and no ... we've posted on the Internet \$15 to \$20 million worth of research and you can access it at no charge. We've just taken our ... I know I'm going to get in trouble with Primary Research for saying this. We just put all of the MarketingSherpa case studies at your disposal, at full availability. We are an institution, not a consultancy, and yes, we have a small part of our business, but we are focused on training and teaching you. And, finally, we have somebody that's actually sponsoring the clinics. It's HubSpot, so be nice to those people. All of this data that's coming to you today and all of this teaching that's coming today is being enabled by the first time sponsorship that we've received. Alright, so let me teach the final point as carefully as I can. Let's begin.

**Selection Effect:** The effect on a dependent variable, by an extraneous variable associated with different types of subjects not being evenly distributed between experimental treatments.

**Plain English Definition:** Selection effect occurs when we wrongly assume some portion of the traffic represents the totality of the traffic.

**Dr. Flint McGlaughlin:** Here is one of those fancy definitions. The effect on a dependent variable by an extraneous variable associated with different types of subjects not being evenly distributed between experimental treatments, or experiential treatments. Now, I read it fast. You understand that it's again a technical definition. And, if you're being certified by us, you'd have to know that. Other than that, it's very useful at parties. When you're going to meet people, you can use phrases like this to impress them with your skills and background, but this is what you need to know. Selection effect occurs when you wrongly or mistakenly assume some portion of the traffic represents the totality of the traffic. Big problem! Often, it's a big problem. In fact, many times we run our tests with our best list, not realizing that our best list is not our best representation of our overall traffic. And, we will get a yield and a result, and an exciting thing to report that doesn't translate when we push it all the way across the site, because our best list, our house list, our best email lists are highly motivated. They have greater levels of trust for us. Many of them are previous customers and they don't represent the marketplace that we're really trying to reach with a new offering, and so be aware.

### Validity Threat #3: Selection effect

---



**Experiment ID:** *Protected*

**Location:** MarketingExperiments Research Library

**Test Protocol Number:** TP2047

#### Research Notes:

**Background:** An ecommerce site focusing on special occasion gifts

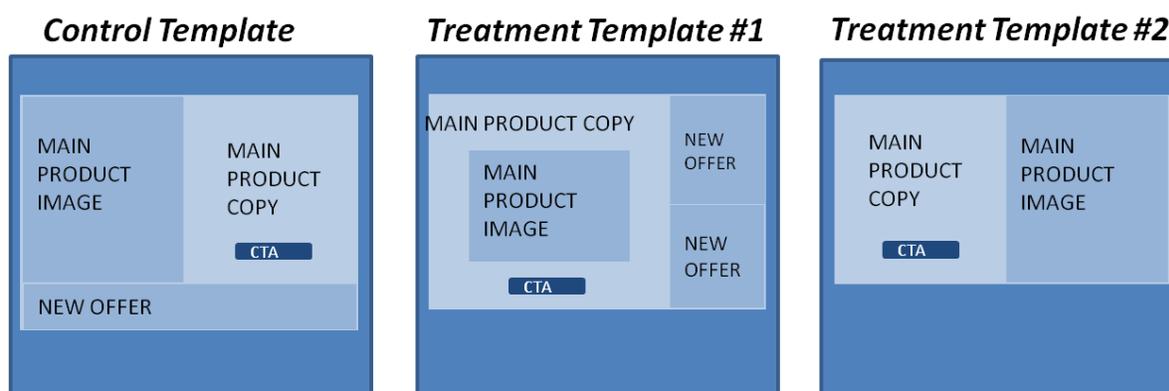
**Goal:** To increase clickthrough and conversion

**Primary research question:** Which email design will yield the highest conversion rate?

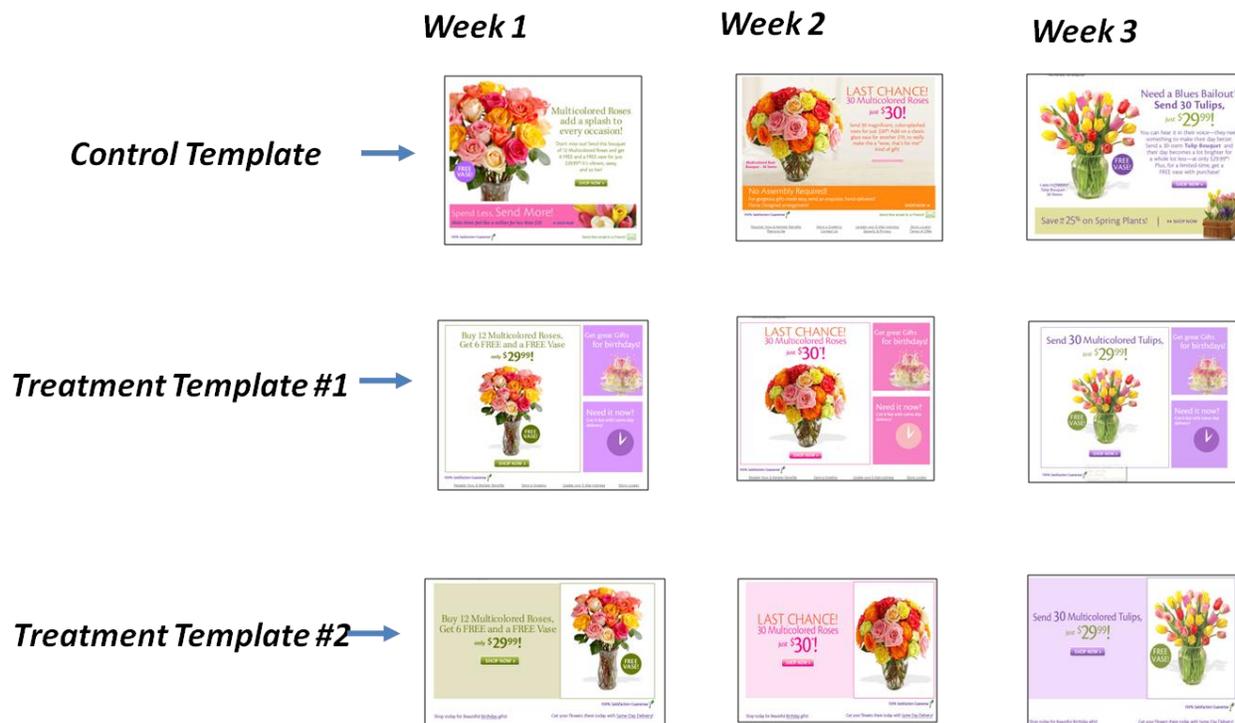
**Approach:** Series of sequential A/B variable cluster split tests

**Dr. Flint McGlaughlin:** Here's an example. Here's an e-commerce site focusing on special occasion gifts. This is Test Protocol 2,047. We've tested about 10,000 of these paths and this one was to increase clickthroughs and conversion, and the question was which email design will yield the highest conversion rate.

In a series of tests lasting 5 weeks, we tested 7 different email templates designed for their most loyal customer segment. Below are examples of three of those email templates tested.



**Dr. Flint McGlaughlin:** And, so here's the control template, here is the Treatment Template 1 and Treatment Template 2. This is a series of tests lasting five weeks. We tested seven different email templates, and it was designed to test their most loyal customer segment. Below, are examples of three: one, two, three. You've got them. Let's keep going.



**Dr. Flint McLaughlin:** Here is week one, and week two, and week three, and here are the three emails within the series. So, you can see them. This is the sell, nine emails on the screen. And, the top would be your controls. This is what we're trying to beat. Let's continue and look at the data set.

 **74% Increase in Conversion**  
*Simple side-by-side layout outperformed the Control*

Week 1 Results		
Template Version	CR	Rel. Diff.
Control Template	14.01%	-
Treatment Template #1	17.06%	25.68%
<b>Treatment Template #2</b>	<b>24.38%</b>	<b>74.05%</b>

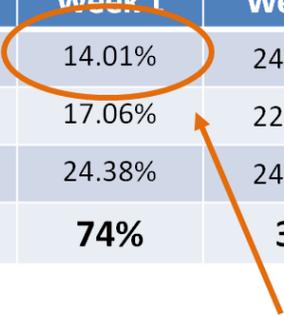


**What you need to understand:** After a week of testing, Treatment 2 converted at a rate 74.05% higher than the control. However, as the subsequent tests were conducted, there was a noticeable shift in results.

**Dr. Flint McGlaughlin:** Success! A 74% increase in conversion. It's the kind of thing that you see regularly if you tune into one of our clinics and we show it to you, and everybody says, "What can we learn from this?" And, the answer is probably not what you're expecting. Sorry to say, we have a data problem somewhere. After a week of testing, Treatment 2 converted a rate of 74% higher than the control. But, notice this, however. I hate the however. However, as the subsequent tests were conducted, there was a noticeable shift in results. Let's look at this.

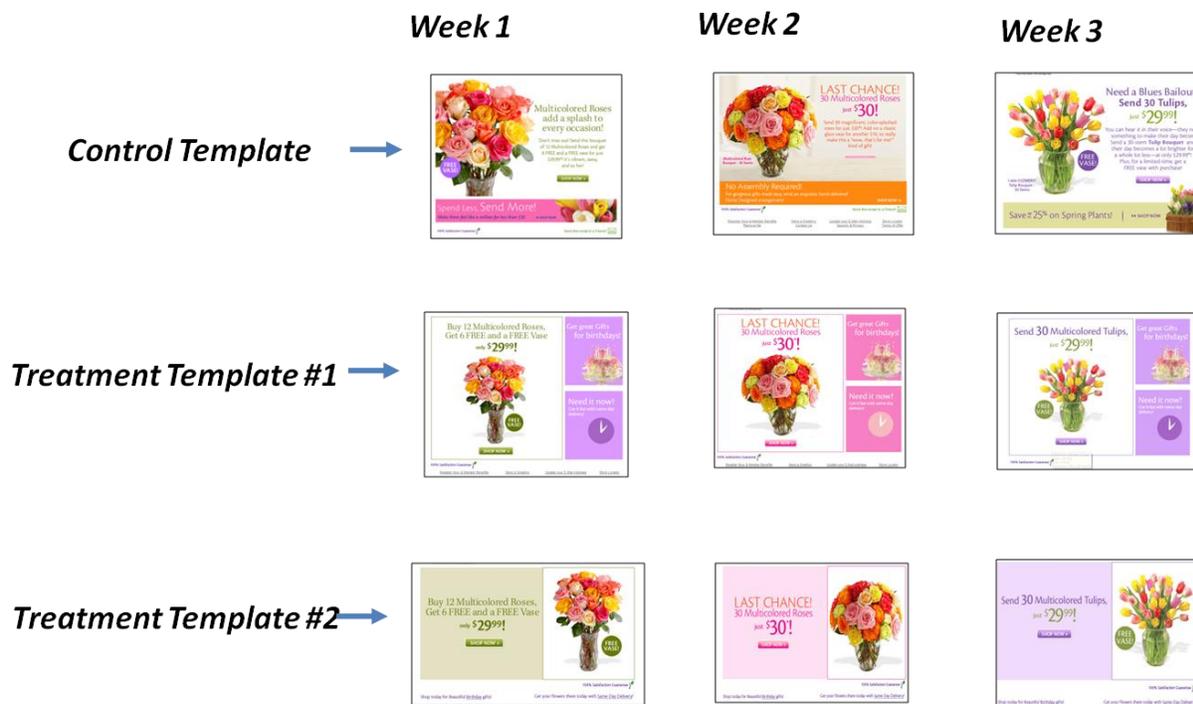
For the remaining test duration, the results never get above 7% indicating the need examine the results and verify whether the test was actually valid.

	Week 1	Week 2	Week 3	Week 4	Week 5
Control	14.01%	24.08%	19.04%	19.77%	20.05%
Treatment #1	17.06%	22.59%	19.09%	19.42%	19.52%
Treatment #2	24.38%	24.89%	20.74%	17.93%	20.50%
<b>Rel. Diff. (T2)</b>	<b>74%</b>	<b>3%</b>	<b>7%</b>	<b>-9%</b>	<b>2%</b>



As we drilled down into the numbers, we learned that it was the selection effect concerning the distributed traffic directed to the control of week 1.

**Dr. Flint McGlaughlin:** For the remaining test duration, the results never got above 7%, indicating something is wrong. And, as we drilled down into the numbers, we learned that it was the selection effect concerning the distributed traffic directed to the control of week one. What am I saying here? Let's go back.



Dr. Flint McLaughlin: So, we have essentially three paths being tested, and those paths are represented with these email designs. And, we look like we're getting a 74% increase.

**74% Increase in Conversion**  
*Simple side-by-side layout outperformed the Control*

Week 1 Results		
Template Version	CR	Rel. Diff.
Control Template	14.01%	-
Treatment Template #1	17.06%	25.68%
<b>Treatment Template #2</b>	<b>24.38%</b>	<b>74.05%</b>

 **What you need to understand:** After a week of testing, Treatment 2 converted at a rate 74.05% higher than the control. However, as the subsequent tests were conducted, there was a noticeable shift in results.

**Dr. Flint McGlaughlin:** Notice CR. I wish you were in the classroom with me right now, but 14.01%, 17.06%, 24.38%, this is the conversion rate. That's the column you want to look at as you go to the next one. I'm flipping to the next column.

In the subsequent weeks of testing these email templates, the overall results of the treatment templates declined to as low as -6% for Treatment 1, and as low as 6% for Treatment 2.

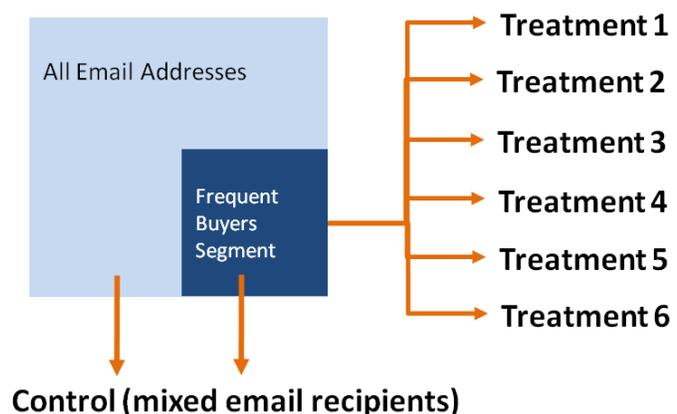
Week 2 Results		
Template Version	CR	Rel. Diff.
Control Template	24.08%	-
Treatment Template #1	22.59%	-6.17%
Treatment Template #2	24.89%	3.38%

Week 3 Results		
Template Version	CR	Rel. Diff.
Control Template	19.04%	-
Treatment Template #1	19.09%	-1.60%
Treatment Template #2	20.74%	6.89%

**Dr. Flint McGlaughlin:** Now, we start to see a problem. Above, you can see that the differences are not that high in week two. And, in week three, so 24, 22 and 24, see how tightly, see how close they are? Look in the next week: 19, 19 and 20. Now, if I were to go back again, 14, 17 and 24 have big differences. Week two does not look this way. Week three does not look this way. And, so something is wrong with our numbers.

During the first week, the treatments received evenly distributed traffic coming from a specific segment of frequent buyers

However, the control received traffic from a mixture of their frequent buyers and the rest of their email list



**Dr. Flint McGlaughlin:** Here's what's going on. During the first week, the treatments received evenly distributed traffic coming from a specific segment of frequent buyers. However, the control received traffic from a mixture of their frequent buyers and their general email list. Are you understanding? This is like Sesame Street. One of these things is not like the other. What's happening is the control didn't get a fair shake. The control got traffic that was mixed between highly loyal, highly motivated, and the general flow of traffic into their site. What was the difference? Well, as soon as that traffic leveled off and they all got three ... all three paths got the same kind of traffic, we didn't see the big difference, we didn't see the big win, and we had nothing we could brag about, just a test that taught us something very important. And, remember, the goal of the test is to get a learning, not a lift. The lifts will come if you get the learnings right.



### 74% Increase in Conversion

*Treatment 2 converted 74.05% more recipients than the Control*

Week 1 Results		
Template Version	CR	Rel. Diff.
Control Template	14.01%	-
Treatment Template #1	17.06%	25.68%
<b>Treatment Template #2</b>	<b>24.38%</b>	<b>74.05%</b>



**What you need to understand:** The difference in the distribution of email recipients between the control and treatments caused enough of a validity threat within the first week that the data has to be excluded from analysis.

**Dr. Flint McGlaughlin:** So, let's go back here for just a moment and look at that test. This is what we really had. We had a data flaw. We had a problem and we had to start over again. But, we learned something very important about parsing that traffic.

## Audience Question

---



Is there a "simple" calculation to use, to know when a test is statistically significant?  
- Several Clinic Registrants

**Dr. Flint McGlaughlin:** So, let's point you to something. I'm just skipping right ahead. I want to help you. Here's a tool. You can download this. We think we already have you on our email list, so this is not our attempt to capture email addresses. This is us just trying to help you out with tool that will help you calculate statistical significance. You can download it right here at this link, [MarketingExperiments.com/ValidityTool](http://MarketingExperiments.com/ValidityTool).

And onward we go. Just get the tool! It'll help you!

## Summary: Putting it all together

---



### Key Principles

1. Just because a test *looks* conclusive doesn't mean it's conclusive.
2. There are at least 3 validity threats beyond sample size that you need to consider when testing:
  - History effect
  - Instrumentation effect
  - Selection effect

**Dr. Flint McGlaughlin:** Moreover, here is a key, here are the three effects you need to take back to your marketing team right now and need to say, "Look, this is what I learned today and these are the three we've got to watch for." Moreover, you can get a copy of this clinic. Pretty soon we'll be releasing it online. And, you can share it. You can sit down with your team and watch it again, same audio, with the same slides. And, you guys can integrate this into your marketing culture.

I want to end here. I want to thank you. If you enjoyed today, there's really one thing you could do for us that would make a great difference here, and that is tell someone about these clinics. We hold them once or twice a month, every month, releasing the latest experiments, and briefings, and discoveries. We've been doing them for years. And, we're trying to build and have been pleased to discover that we can aggregate a huge community of marketers who are helping each other figure out what really works. That's our mission. Thank you!

# MarketingExperiments Journal

**FREE subscription to more than \$10 million in marketing research**

Join 98,000 of the top marketers from around the world as we work together to discover what really works.

View with Images | View Mobile Version



powered by MECLABS

[Free Subscription](#) | [Research Journal](#) | [Methodology](#) | [Research Directory](#) | [Training](#)

Welcome to the MarketingExperiments Journal email for March 17, 2011. Every three weeks we send this high-level update of our research to 98,000 marketers. Today you'll find:

- The on-demand replay of [Do You Have the Right Value Proposition?](#), where we taught five simple steps to determine your optimal value proposition
- An invitation to our next Web clinic, where we will discuss proven strategies for [Converting Leads to Sales](#)
- An opportunity to [share your optimization and testing discoveries](#) with your peers at the 2011 Optimization Summit and receive a complimentary ticket to this inaugural event.

Our job is to help you do your job better. [Let us know](#) how we can help.

Daniel Burstein  
Director of Editorial Content  
MECLABS Primary Research

Resource #1

**Do You Have the Right Value Proposition?**  
**How to test, measure, and integrate your Value Proposition online**

According to Dr. Flint McGlaughlin, there are five simple steps any marketer can take to identify their optimal value proposition. During this dense, yet very practical, Web clinic, he reviewed two recent experiments aimed at discovering value propositions that increase customer response.

Connect with us:

- [Visit our blog](#)
- [Visit our website](#)
- [Follow us on Twitter](#)
- [Join us on LinkedIn](#)
- [Forward to a colleague](#)

Latest on our blog:

[Online Testing and Optimization: ROI your test results by considering effect size](#)

[E-commerce: Using multivariate testing to increase sales 83.79%](#)

[Landing Page Optimization: Minimizing bounce rate with clarity](#)

**With your FREE subscription you receive:**

- First access to \$10 million in optimization research
- Four live web clinic invitations per quarter
- One Research Journal per quarter

**Subscribe for FREE!**

MarketingExperiments.com/subscribe

# Discover What Really Works in Optimization

[MarketingExperiments](#) is a primary research facility, wholly-owned by [MECLABS](#), with a simple (but not easy) seven-word mission statement: **To discover what really works in optimization.**

We focus all of our experimentation on optimizing marketing communications. To that end, we test every conceivable approach and we publish the results in the *MarketingExperiments Journal* ([subscribe](#)).

## Three ways to make the most of MarketingExperiments:

1. **Self-Guided Learning:** Access, for *free*, more than \$10 million in primary marketing research and experiments via our [web clinics](#), [blog](#) and [research directory](#).
2. **Formal Training:** Learn how to increase your marketing ROI through [live events and workshops](#), [online certification courses](#) and [live company training](#).
3. **Research Partnership:** Apply for a [research partnership](#) and let the MarketingExperiments team help drive conversions and ROI for your subscription, lead-generation, ecommerce, email and other online marketing efforts

Would you like to learn the MarketingExperiments optimization methodologies from the inside out? We're always looking for the next great optimizer to push our research forward. Learn more on our [careers page](#).

## Share your success and learnings

While we at MarketingExperiments are glad to share what we've discovered about optimization to date through our own experimentation, we also publish case studies and completed tests to facilitate peer-learning from real marketers with real challenges.

To that end, we're always looking to shine a light on your hard work. If you have a success or learning you'd like to share, [let us know](#).