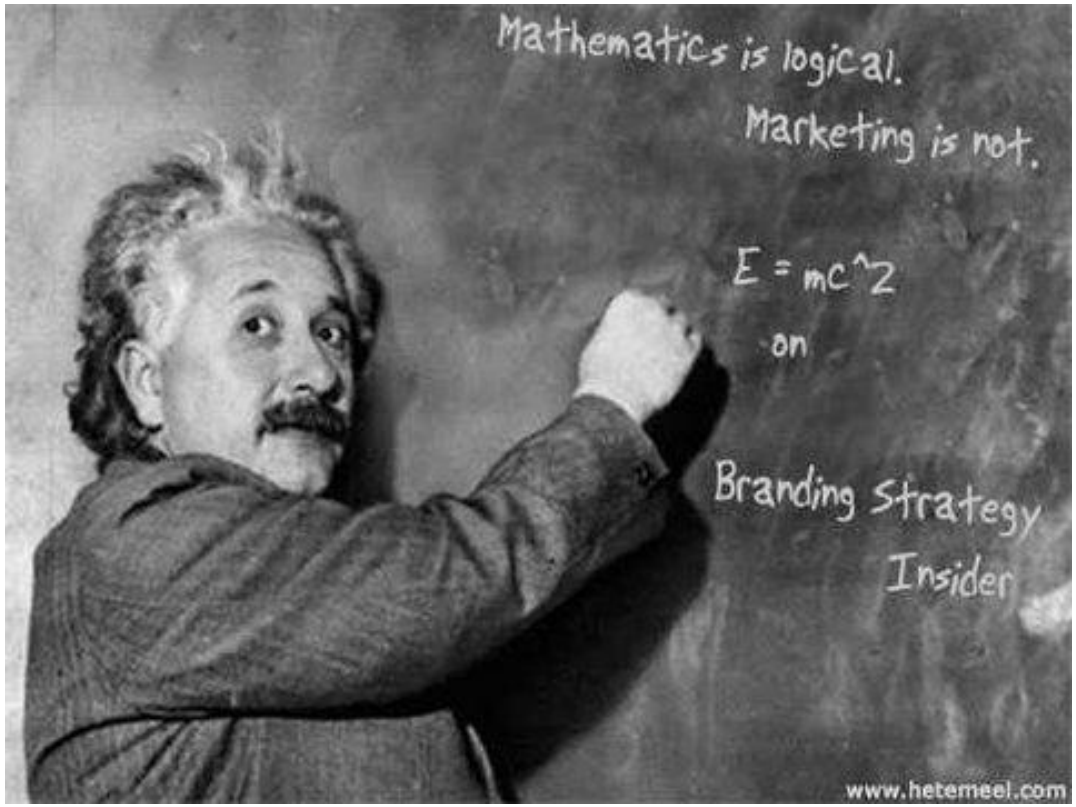
Validity for Decision Makers

A guide to making optimization decisions with confidence

Sergio Balegno
Director of Research
MECLABS

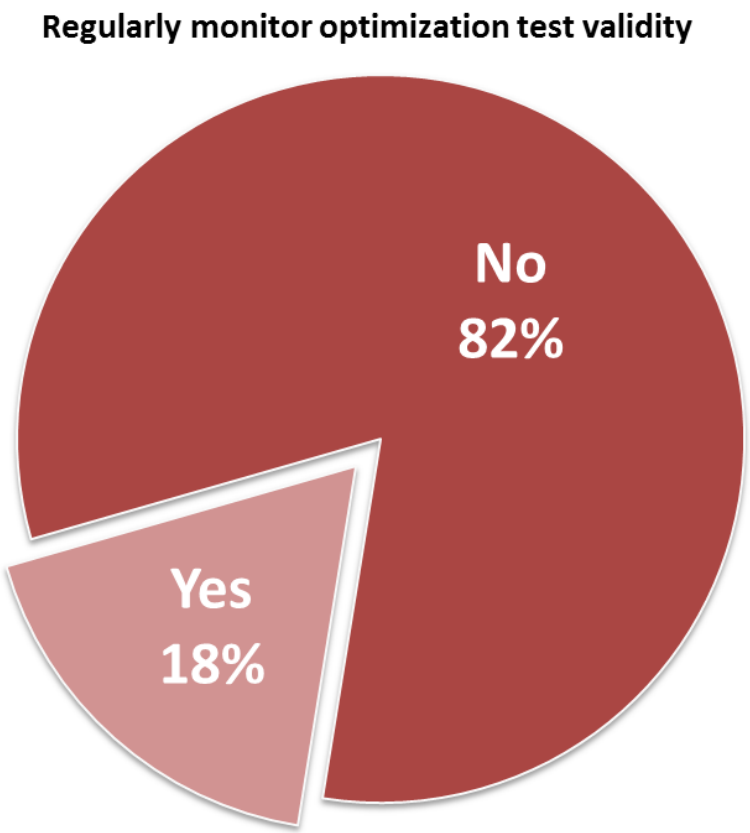
Bob Kemper
Sr. Director of Sciences
MECLABS

What we don't know about validity can hurt us



- Do you monitor the validity of your tests?
- How well do you understand the math used to determine statistical validity?

82% of marketers surveyed do not monitor optimization test validity



marketingsherpa Source: ©2012 MarketingSherpa Website Optimization Benchmark Survey
Methodology: Fielded April 2012, N=2677

- If you don't, then you're in good company…

  … or at least you're surely not alone!

Validity for Decision Makers

1. An invalid test…
   … is worse than no test at all
   Principles and Taxonomy of Optimization Testing
2. Test validity:
   There's more to it than 'Math'
3. Validity Threats
   History Effect
   Instrumentation Effect
   Sample Selection Effect
   Sample Distortion Effect
4. Math and practical examples

Experiment ID: *Protected*
Location: MarketingExperiments Research Library
Test Protocol Number: TP2047

Research Notes:

Background: An ecommerce site focusing on special occasion gifts

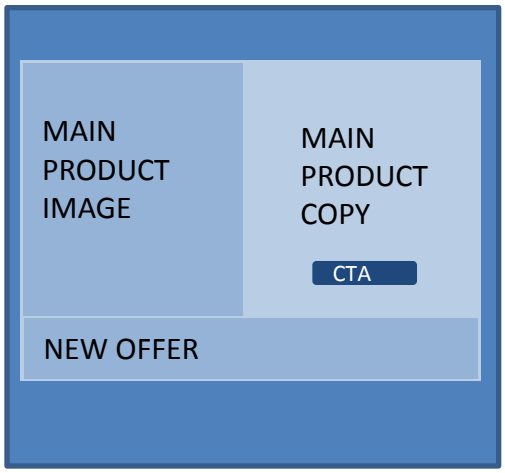
Goal: To increase email clickthrough and conversion-to-sale

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

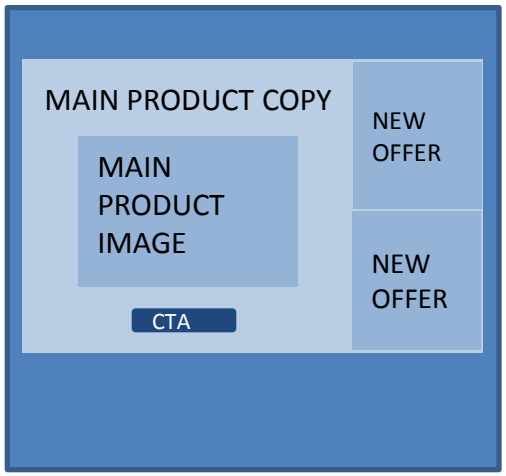
Approach: Series of A/B variable cluster split tests

- In a series of daily email 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.

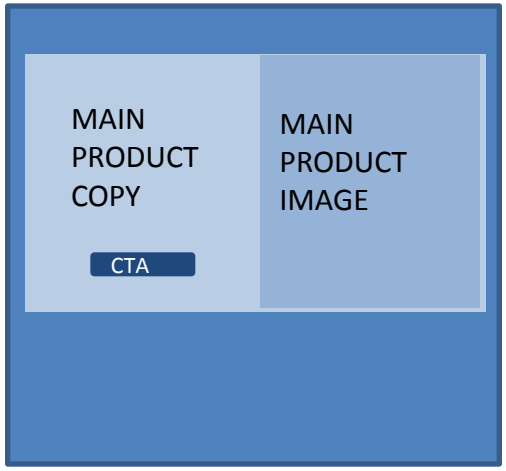
Control Template



Treatment Template #1



Treatment Template #2



...

	Week 1	Week 2	Week 3
Control Template →			
Treatment Template #1 →			
Treatment Template #2 →			

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%
Treatment Template #2	**24.38%**	**74.05%**

What you need to understand: After a week of testing, Treatment 2 converted at a rate 74.05% higher than the control.

However, as subsequent samples were collected, there was a noticeable shift in results.

- In the subsequent 2weeks, the relative conversion rates of the Experimental Treatment templates declined to as low as -6% for Treatment 1, and as low as 3% for Treatment 2 compared to Control.

Week 2 Results

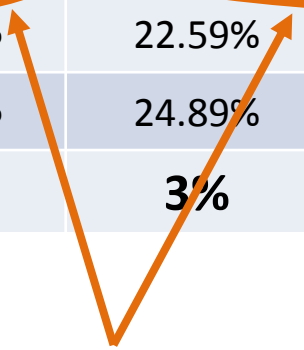
Template Version	CR	Rel. Diff.
Control Template	24.08%	-
Treatment Template #1	22.59%	-6.17%
Treatment Template #2	24.89%	3.38%

Week 3 Results

Template Version	CR	Rel. Diff.
Control Template	19.04%	-
Treatment Template #1	19.09%	0.26%
Treatment Template #2	20.74%	8.93%

- For the remainder of the 5 week test, the relative CR never again exceeded +9%, but stabilized indicating that something had potentially invalidated the first week of tests.

	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%
Rel. Diff. (T2)	**74%**	**3%**	**9%**	**-9%**	**2%**



- As we drilled down into the numbers, we learned that there was a problem during the first week related to incoming traffic to the Control.

We'll come back to this test a little later, when we talk about specific validity threats…

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%
Treatment Template #2	**24.38%**	**74.05%**

☑ **What you need to understand**: The apparent significant performance boost indicated after Week 1 was in fact a 'phantom' gain.

There actually was no statistically significant difference among the treatments.

An invalid test…
… is worse than no test at all.

Those who neglect to test know the risk they're taking, and make their changes cautiously and with a healthy trepidation.

Those who conduct invalid tests are blind to the risk they take, and make their changes boldly and with an unhealthy sense of confidence.

Validity for Decision Makers

The purpose and taxonomy of optimization testing

☞ *Key Principle*

In online testing, the ultimate measure of a test is its **Utility**:

In marketing optimization…

A **useful** test is one that helps you predict future customer behavior.

Five elements compose the Utility of an experimental test. For us, they are expressed as…

Online Testing Heuristic:

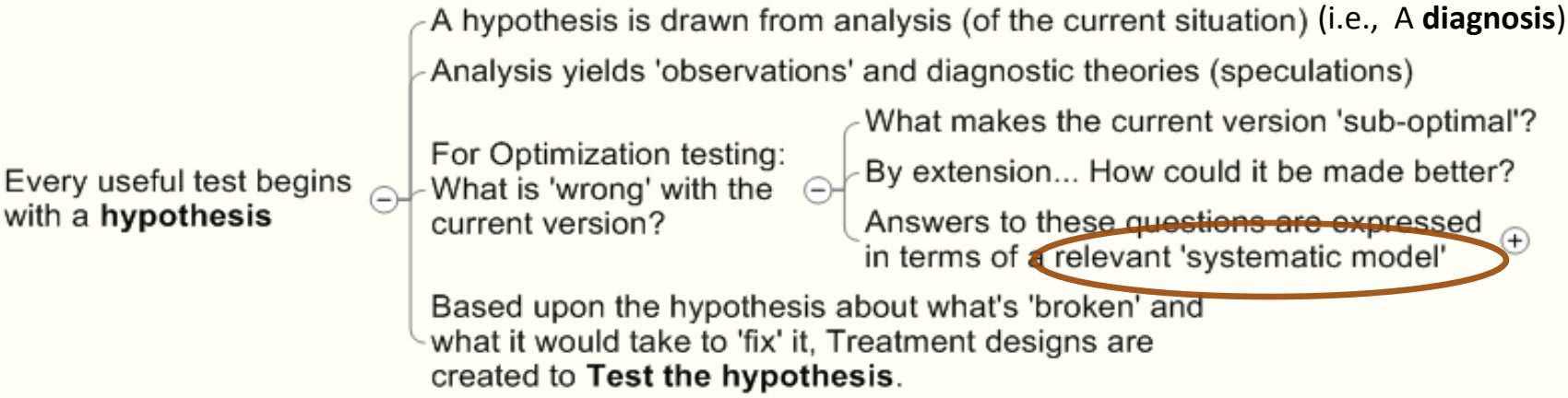
$$u = 2q + t + m + 2v + i^{©}$$

u = Utility
q = Research Question
t = Treatment
m = Metric System
v = Validity Factor
i = Interpretation

☞ *Key Principle*

In online testing, the ultimate measure of a test is its **Utility**:

In marketing optimization…



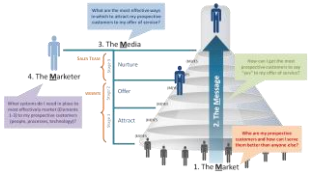
The hypothesis is one that proposes a cause or a **reason that customers are currently behaving** as they do…

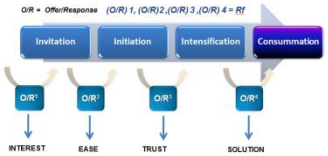
… and describes **how that behavior would differ** if something specific were altered about the experience they encounter (when they receive your email, when they arrive at your site, …)

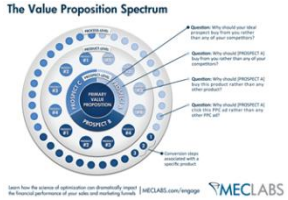
Process: **Diagnosis → Hypothesis → Treatment Design & Development**

Systematic Models

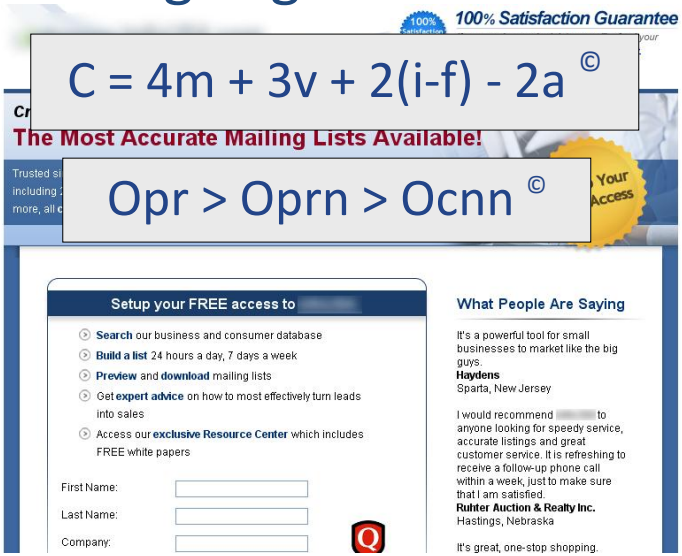
Overarching Meta-theory







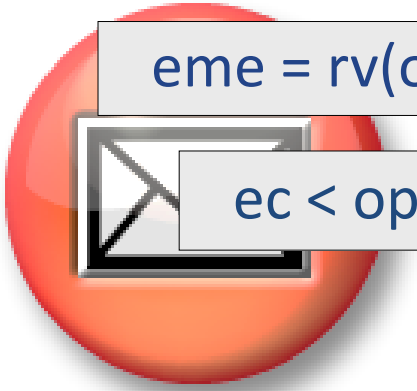
Landing Pages:



$C = 4m + 3v + 2(i-f) - 2a$ ©

Opr > Oprn > Ocnn ©

Email effectiveness…



$eme = rv(of+ i) – (f + a)$ ©

ec < op < ct < lp ©

☞ *Key Principles*

- **Ultimately, a test is designed for one purpose…**

  To help us **"Accept"** or **"Reject"** a **hypothesis.**

  - **The extent to which it does this is the measure of its <u>Utility</u>**:

☞ Key Principles

- **Ultimately, a test is designed for one purpose…**

  To help us **"Accept"** or **"Reject"** a **hypothesis.**

  - **The extent to which it does this is the measure of its Utility:**

- **Experimental scientists work (and think) backward…**
  - "Normal" people seek to use testing and experimental math to 'prove' a new page converts better than the old one.
  - Experimental scientists seek to use probability and statistics to show that it's extraordinarily unlikely that the two pages are really 'the same'.
    (Their greatest joy, in fact is the opportunity to do something (uninspiringly) termed "rejecting the Null hypothesis".) ….zzzzzzzzzzzzzzzzz…
    - We'll talk about the reason for this when we get to the 'Math' part of validity

Validity for Decision Makers

Validity Threats

Key Principles

In online testing, there are four common validity threats that must be considered:

- History Effect
- Instrumentation Effect
- Selection Effect
- Sampling Distortion Effect

Essential Definitions

History Effect – when a test variable is affected by the passage of time

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 click-though?

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

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.

Child Predator Registry
Identify sex offenders living in your area. Protect your kids today.
www.XXXXXXXXXXXXXXXX.com

Is Your Child Safe?
Identify sex offenders living in your area. Protect your kids today.
www.XXXXXXXXXXXXXXXX.com

Predators in Your Area
Identify sex offenders living in your area. Protect your kids today.
www.XXXXXXXXXXXXXXXX.com

Find Child Predators
Identify sex offenders living in your area. Protect your kids today.
www.XXXXXXXXXXXXXXXX.com

- During the test, Dateline aired a special called "*To Catch a Predator*" viewed by approximately 10 million individuals
- Throughout this program sex offenders are referred to as "predators"

Headline	Impressions	Clicks	CTR
Predators in Your Area	21,096	1,423	**6.74%**
Child ***Predator*** Registry	14,712	652	**4.43%**
Find Child ***Predators***	18,459	817	**4.42%**
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.**

Key Principles

1. In online testing, there are four common validity threats that must be addressed: History Effect, Instrumentation Effect, Selection Effect, and Sampling Distortion Effect.

2. To minimize the **History Effect**, monitor external events, compare current with historical data, and look for anomalies in the testing data.

Examples of external events that commonly impact test results:

- *Media events (news and entertainment)*
- *Competitor marketing initiatives*
- *Internal marketing initiatives*
- *Economic changes (esp. 'shocks')*
- *Political changes*

Essential Definitions

History Effect – when a test variable is affected by the passage of time

Instrumentation Effect – when a test variable is affected by a change in the measurement instrument

Experiment ID: *(Protected)*
Location: MarketingExperiments Research Library

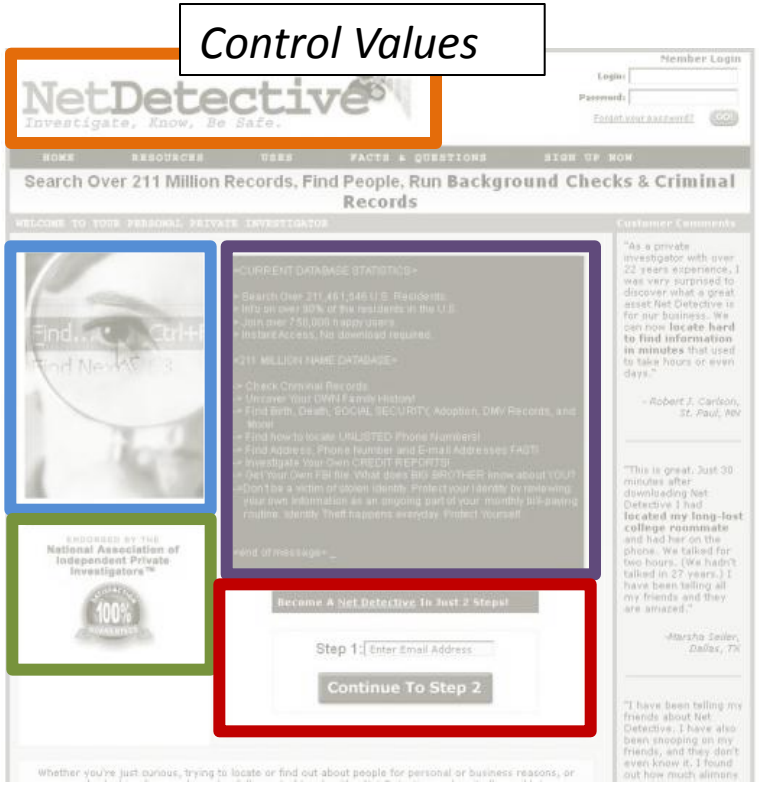
Research Notes:

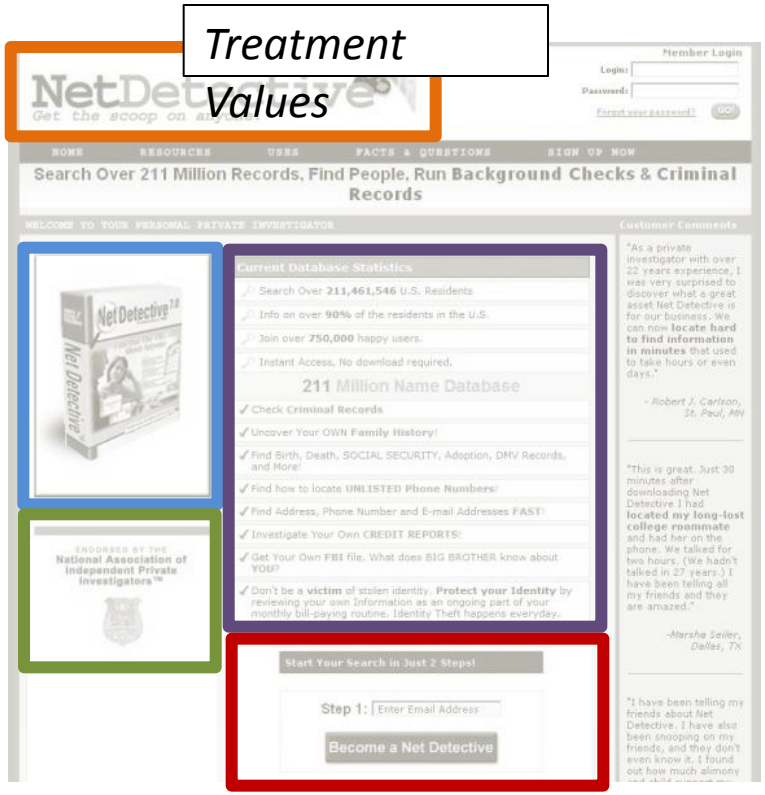
Background: Online 'people search' service offering background checks about criminal records and other public information about individuals, limited to the U.S.

Goal: To increase the rate of Clickthrough to the order-start page, and ultimately to final conversion to sale.

Primary research question: Which landing page will produce the highest Order-start page click-though?

Test Design: Multi-factor (MV) test with 5 factors at 2 levels each.





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

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 versions of experimental variables whenever the Treatment was loaded.
    - This caused the Treatment pages to have substantially longer load times than the Control when they were rendered on the visitor's browser.

Control Page

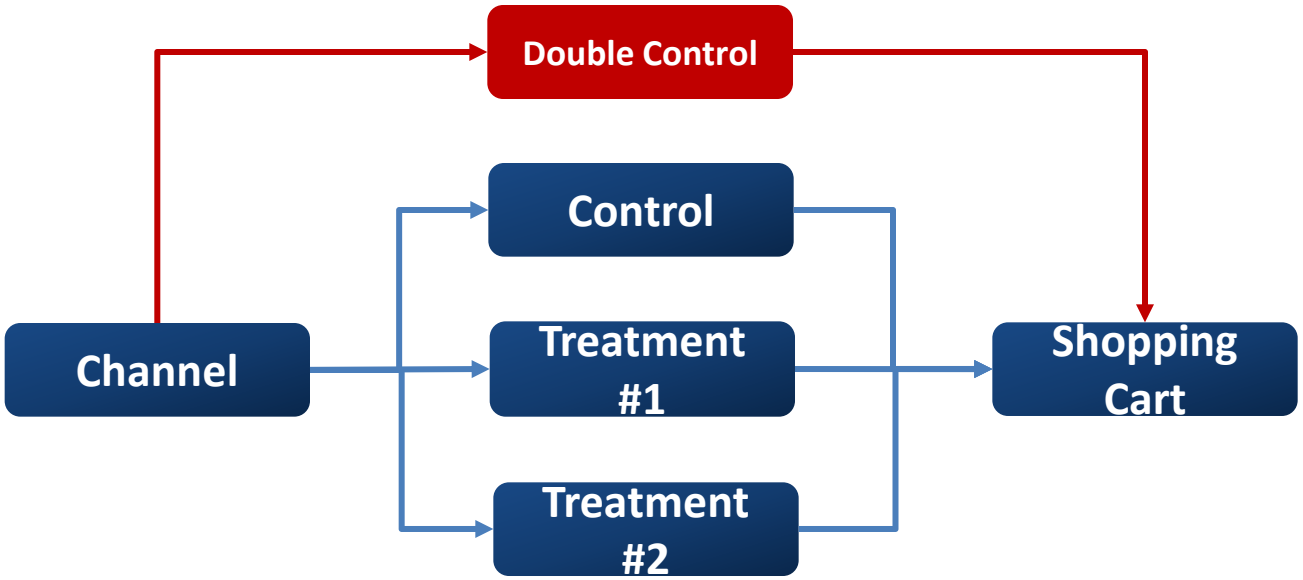
Treatment Page

Page Load Time Reference Chart (in seconds)						
Page Size(kb)	**Connection Rate**					
	14.4	**28.8**	**33.6**	**56**	**128 (ISDN)**	**1440 (T1)**
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
84.9	**60.61**	**30.30**	**25.98**	**15.59**	**6.82**	**0.61**
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
137	**97.80**	**48.90**	**41.92**	**25.15**	**11.00**	**0.98**
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)	**37.19**	**18.60**	**15.94**	**9.56**	**4.18**	**0.37**

☑ **What you need to understand**: The Treatment page took 60% longer to render than Control. At speeds slower than T1, this translated to several seconds. The difference in load times caused the "user experience" to be asymmetric among the page versions, thereby threatening test validity.

Key Principles

1. In online testing, there are four common validity threats that must be addressed: History Effect, Instrumentation Effect, Selection Effect, and Sampling Distortion Effect.
2. To minimize the **History Effect**, monitor external events, compare current with historical data, and look for anomalies in the testing data.
3. To address the **Instrumentation Effect**, never presume your analytics system is accurate; carefully monitor and test for anomalies.



- An additional safeguard against instrumentation anomalies is to include a 'double control' treatment that is identical to the Control for comparison.
- A high variance between Control and Double Control indicates potential problems with instrumentation.
- Note: This approach innately increases test duration

Essential Definitions

History Effect – when a test variable is affected by an extraneous variable associated with the passage of time

Instrumentation Effect – when a test variable is affected by a change in the measurement instrument

Selection Effect – when a test variable is affected by different types of subjects not being evenly distributed among experimental treatments

Experiment ID: *Protected*
Location: MarketingExperiments Research Library
Test Protocol Number: TP2047

Research Notes:

Background: An ecommerce site focusing on special occasion gifts

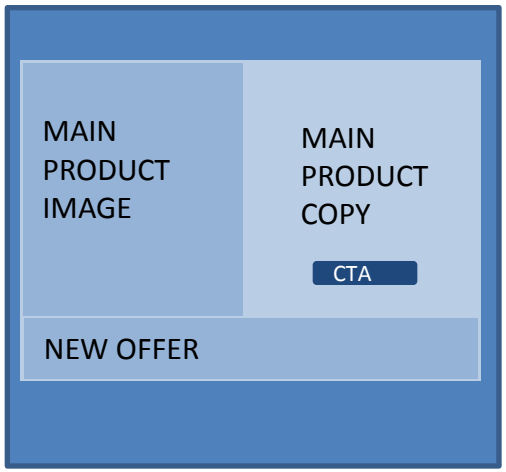
Goal: To increase email clickthrough and conversion-to-sale

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

Approach: Series of A/B variable cluster split tests

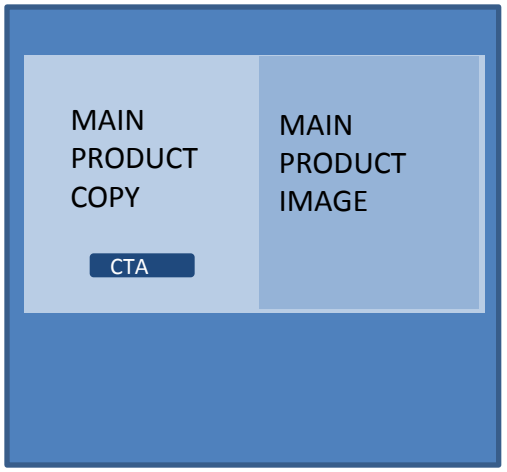
- In a series of tests lasting 5 weeks, we tested 7 different email templates designed for their most loyal customer segment. Below are examples of three of those email templates tested.

Control Template



Treatment Template #1

Treatment Template #2



	Week 1	Week 2	Week 3
Control Template			
Treatment Template #1			
Treatment Template #2			

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%
Treatment Template #2	**24.38%**	**74.05%**

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 test samples continued during the 2nd and subsequent weeks, there was a noticeable shift in results.

- In the subsequent 2weeks, the relative conversion rates of the Experimental Treatment templates declined to as low as -6% for Treatment 1, and as low as 3% for Treatment 2 compared to Control.

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

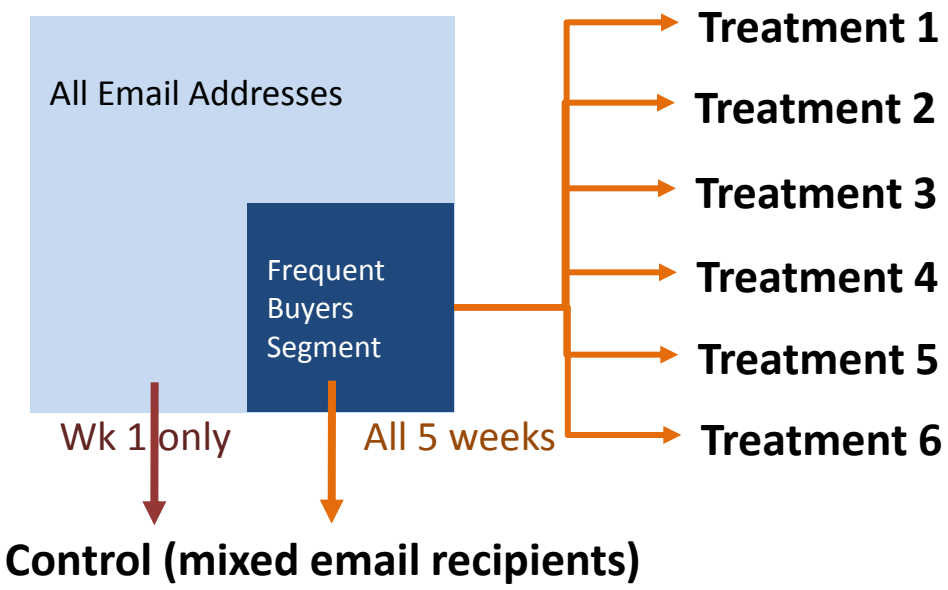
Template Version	CR	Rel. Diff.
Control Template	19.04%	-
Treatment Template #1	19.09%	0.26%
Treatment Template #2	20.74%	8.93%

- For the remaining test duration, the relative CR never exceeded +9%, but stabilized indicating that something has potentially invalidated the first week of tests.

	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%
Rel. Diff. (T2)	**74%**	**3%**	**9%**	**-9%**	**2%**

- As we drilled down into the numbers, we learned that there was a problem during the first week related to incoming traffic to the Control…

- During the entire period, the Treatments and the Control both received evenly distributed traffic coming from a specific segment consisting of their most frequent buyers.
- However, during the first week only, the Control also received traffic from a mixture of the rest of their email list.



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%
Treatment Template #2	**24.38%**	**74.05%**

☑ **What you need to understand**: The systematic difference in distribution of arriving traffic among the treatments caused a Selection Effect validity threat which invalidated the first week's test results. So, the apparent significant performance boost indicated after Week 1 was in fact a 'phantom' gain, and there was no 'real' performance difference among the Control and the Experimental treatments.

Key Principles

1. In online testing, there are four common validity threats that must be addressed: History Effect, Instrumentation Effect, Selection Effect, and Sampling Distortion Effect.
2. To minimize the **History Effect**, monitor external events, compare current with historical data, and look for anomalies in the testing data.
3. To address the **Instrumentation Effect**, never assume your analytics system is accurate and carefully monitor and test for any anomalies.
4. To address the **Selection Effect,** ensure the profile of your test subjects matches as closely as possible the profile of your visitors, and that the assignment of treatments to test subjects is random.

Essential Definitions

History Effect – when a test variable is affected by an extraneous variable associated with the passage of time

Instrumentation Effect – when a test variable is affected by a change in the measurement instrument

Selection Effect – when a test variable is affected by different types of subjects not being evenly distributed among experimental treatments

Sample Distortion Effect – the effect on a test outcome caused by failing to collect a sufficient number of observations

Experiment ID: *(Protected)*
Location: MarketingExperiments Research Library

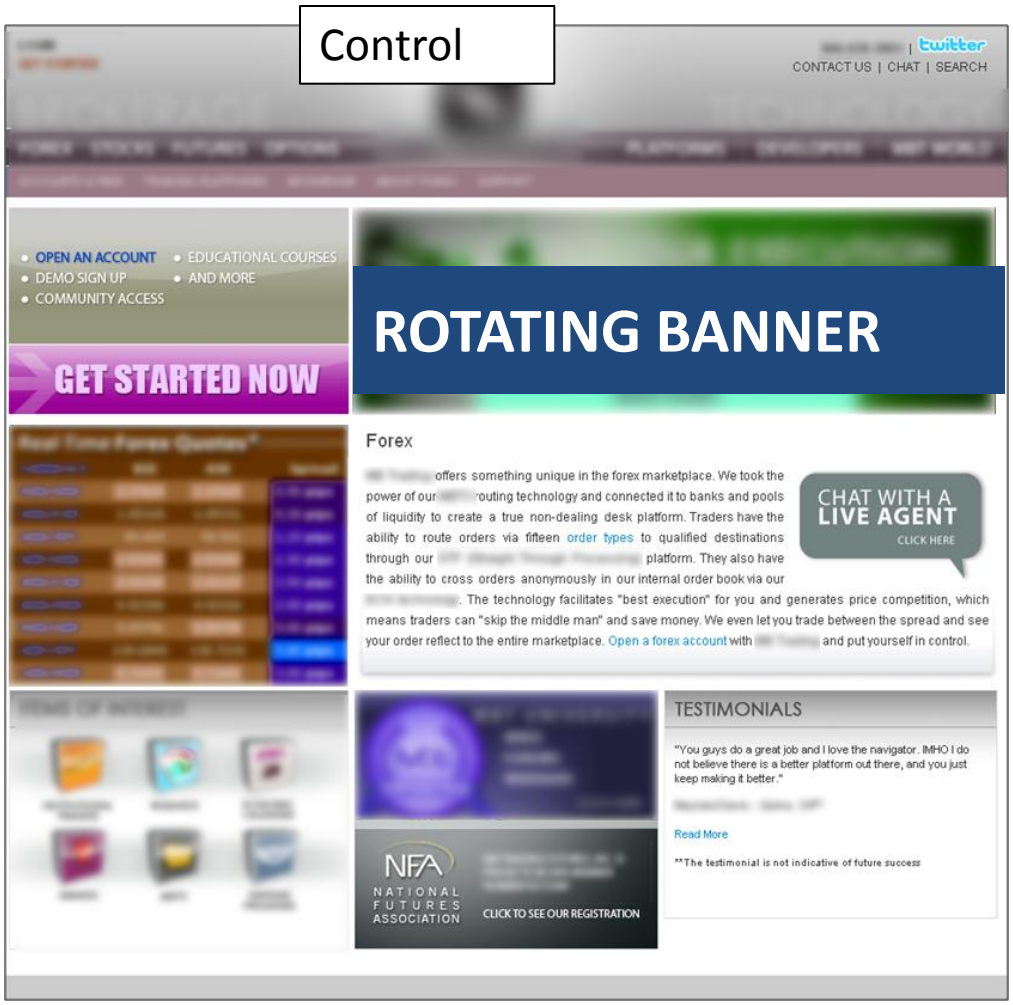
Research Notes:

Background: Consumer company that offers online brokerage services

Goal: To increase the number of accounts created online

Primary research question: Which page design will generate the highest rate of conversion?

Test Design: A/B/C/D multi-factor test



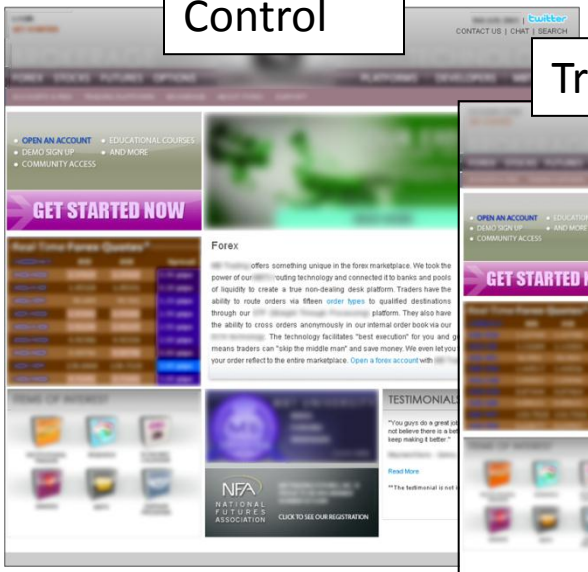
- Heavily competing imagery and messages
- Multiple competing calls-to-action

Validity for Decision Makers – Sample Distortion Effect – Case Study

Hypotheses:

- Difficulty-oriented Friction from multiple competing CTA's is causing drop-out
- Clarity of Value proposition mitigated by lack of visual cues and large block of unbroken text.

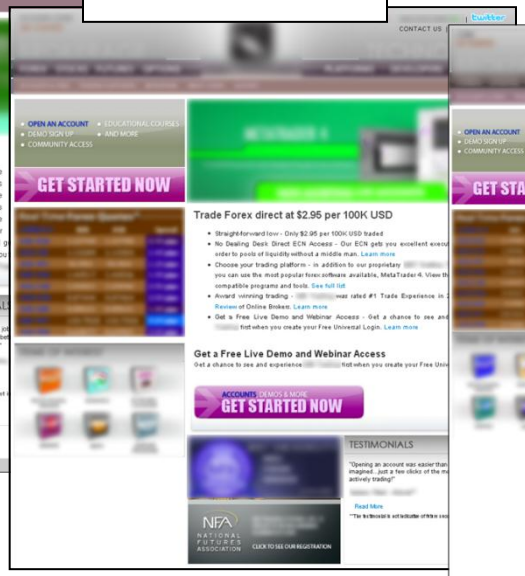
Control



Observations (from CIA)**:**

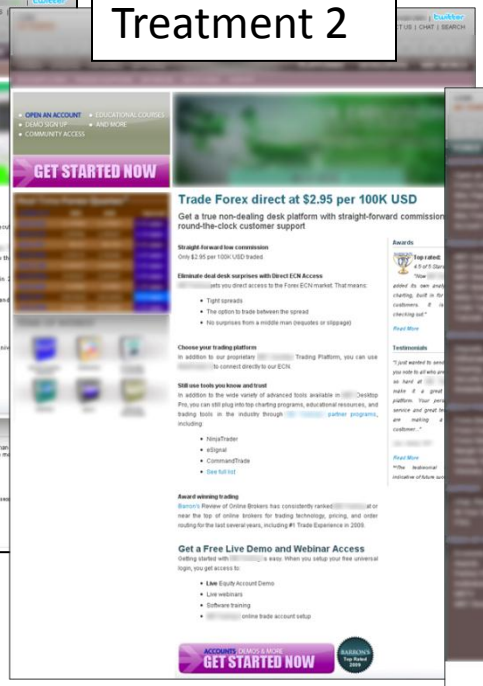
- Heavily competing imagery & messages
- Multiple competing CTA's

Treatment 1



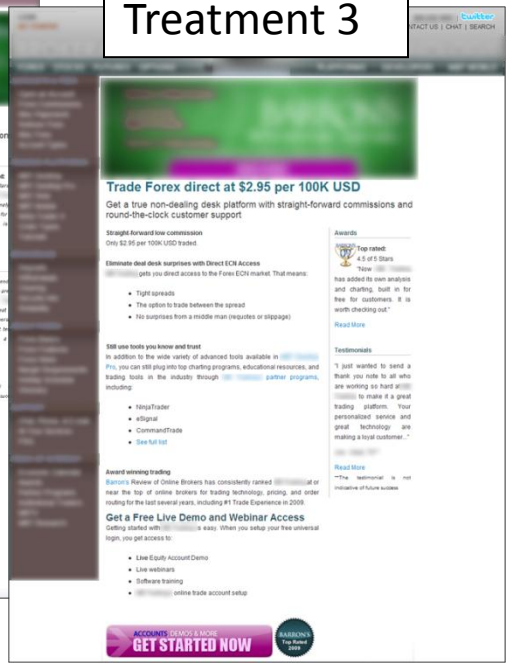
- Most elements on the page remain the same
- Headline added
- Bulleted copy highlights VP items
- Large, clear CTA added

Treatment 2



- Left column is the same, but footer items removed
- Long copy, vertical flow
- Added awards & testimonials in right column
- Large, clear CTA – Like T1

Treatment 3



- Similar to Treatment 2, except left-column reduced even further
- Still a long copy, vertical flow, single call-to-action design

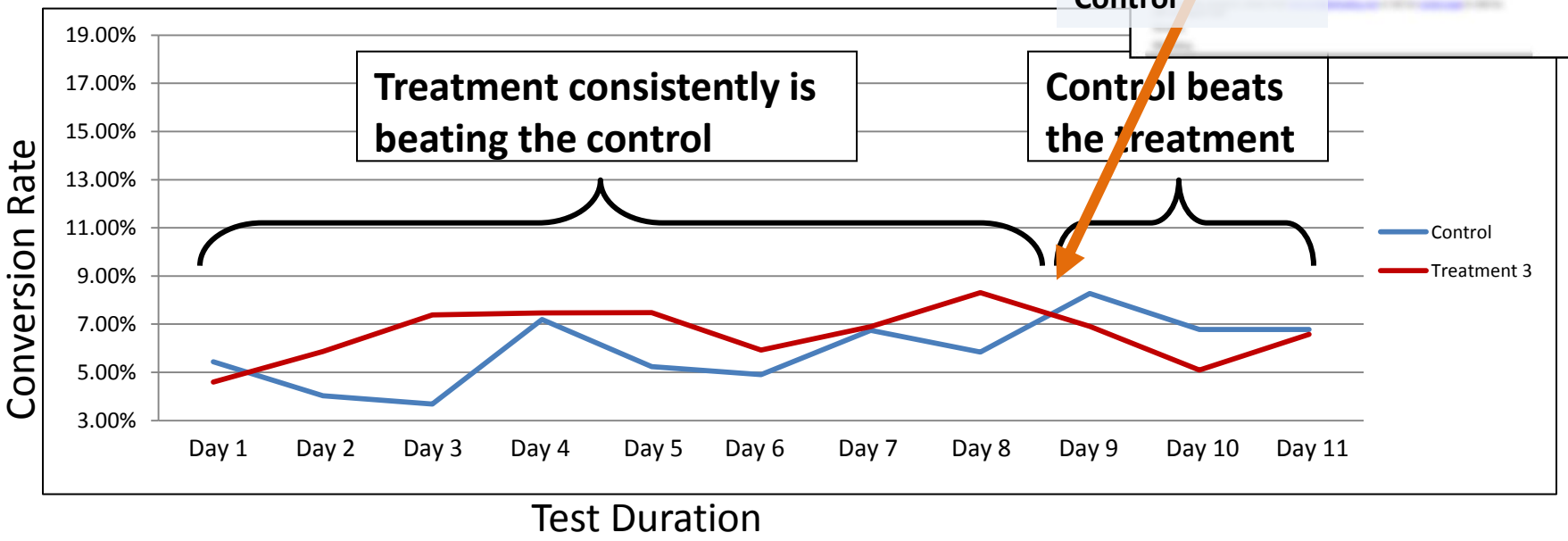
No Significant Variation

None of the Experimental treatments conclusively outperformed the Control

Test Designs	Conversion Rate	Relative Diff.
Control	5.95%	-
Treatment 1	6.99%	17.42%
Treatment 2	6.51%	9.38%
Treatment 3	6.77%	13.70%

What you need to understand: After eleven days, the test management platform reports remained inconclusive. None of the treatments significantly outperformed the control.

- However, a review of the daily sample data showed a distinct shift in relative performance beginning on Day 8 and lasting 2 days.
- Subsequent investigation revealed that during the test, an email was sent that drove traffic to Control, skewing the sampling distribution.



31% Increase in Conversions

The highest performing treatment outperformed the Control by 31%

Test Designs	Treatment	Variation
Control	5.35%	-
Treatment 1	6.66%	25%
Treatment 2	6.13%	15%
Treatment 3	**7.03%**	**31%**

What you need to understand: Inclusion of only the data collected prior to the renegade email send established that Treatment 3 significantly outperformed Control. Measured Level of Confidence: 95.6%.

This underscores the need for a simple way to address the statistical validity threats associated with the testing sample .

Validity for Decision Makers

Validity Threats: Sample Distortion Effect

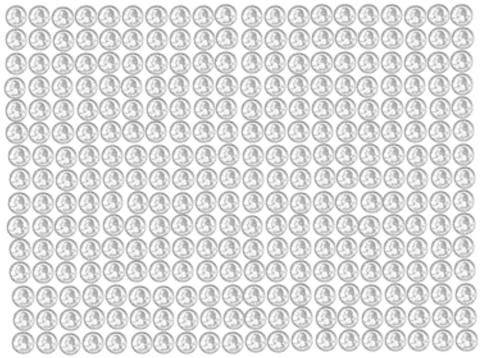
Validity for Decision Makers – Sample Distortion – The Challenge

QUESTION: *At what point could we confidently assume that a coin (possibly a trick coin), when flipped, will land on heads rather than tails?*

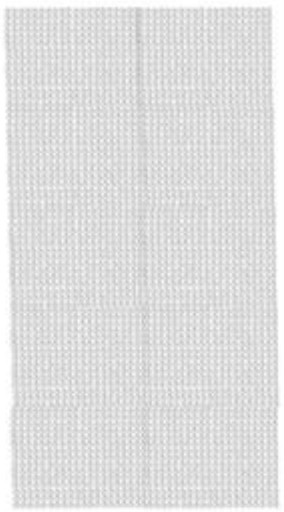
After **3** consecutive flips of heads?

After **30** consecutive flips of heads?

After **300** consecutive flips of heads?



After **3000** consecutive flips of heads?



*For a **Fair** coin…*

p(all H) = 1 in **8**

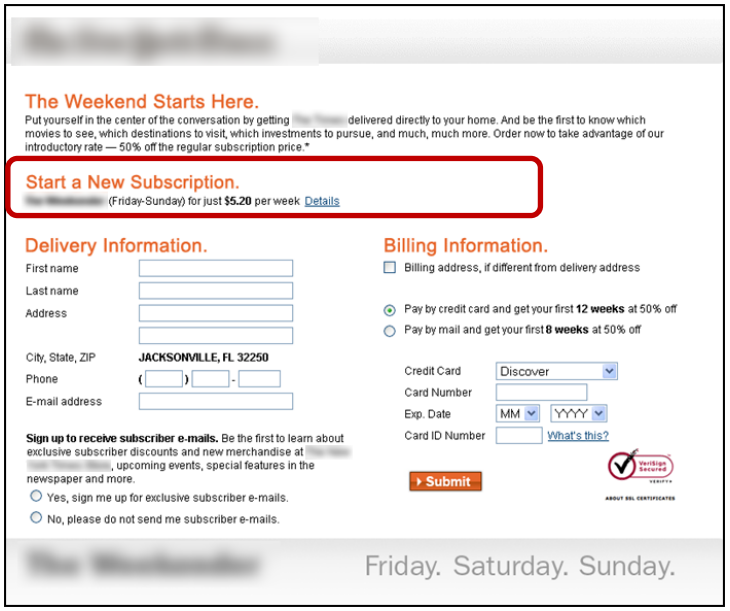
p(all H) = 1 in **1,073,741,824**

p(all H) = 1 in
2,037,035,976,334,490,000,000,000,
000,000,000,000,000,000,000,000,0
00,000,000,000,000,000,000,000,00
0,000,000,000,000,000.
(i.e., 1 in 2.037×10^{90})

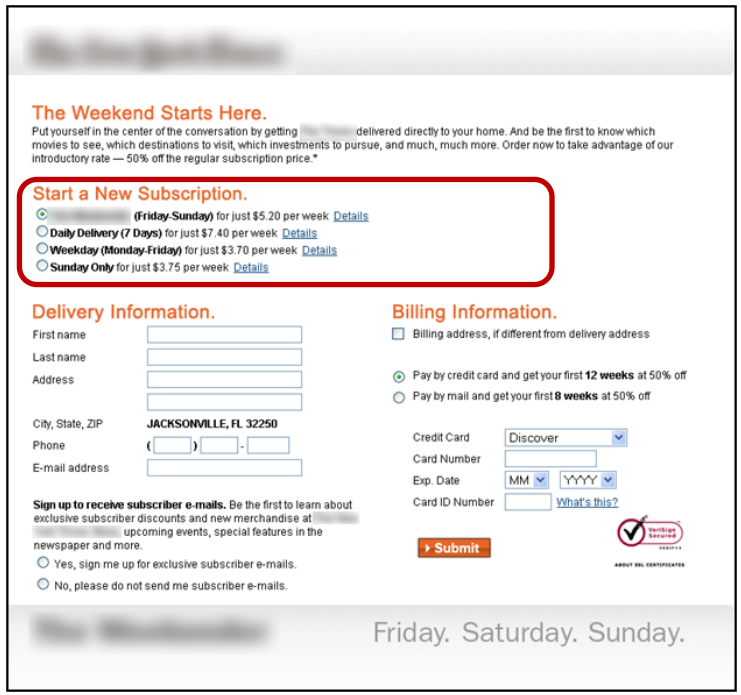
p(all H) = 1 in (**a really big #)**

p(all Heads) = $1/2^n$

QUESTION: *At what point could we confidently assume that one web page will outperform another?*



Vs.



After Three days?
Five days?
Seven days?

Fifty seven days?

Validity for Decision Makers

Statistical Confidence (in Theory)

☞ *Essential Definitions*

Sample Distortion Effect – the effect on a test outcome caused by failing to collect a sufficient number of observations

Statistical Level of Confidence – the statistical probability that there really is a performance difference between the control and experimental treatments based upon the samples collected to date

H_0 is that there is NO difference…

Key Principles

1. There are essentially three factors that determine an outcome's statistical certainty:

   - **The Amount of Difference** – The magnitude of the difference in the primary metric between the Control and an Experimental treatment. (usually expressed as a percentage value)
   - **The Success Rate** – The number of experimental trials in which the outcome was favorable with respect to the primary success metric divided by the total samples received. (usually expressed as a percentage value)
   - **The Samples Received** – The total number of experimental trials used to compute the measured sample success rate. (e.g., number of visitors to a landing page, number of email messages sent).

2. An outcome's true statistical certainty is then compared with the desired level of certainty (chosen prior to experimentation).

THE MATH: *Statistical principles underlying Significance and Level of Confidence as it relates to online optimization testing*

Principle: Distinguish between Random Variation and a 'Real' systematic difference. I.e., what are the chances that what I am observing could happen even though the two treatments were really IDENTICAL?

Mathematical Method:

Determine:

What is the best estimate of the **magnitude of difference** between the observed relative performance of the treatments?

Using: Observed absolute 'Difference in Sample Means'

(Avg %success$_T$) – (Avg % success$_C$)

$$\bar{S} = \frac{\sum s_T}{n_T} - \frac{\sum s_C}{n_C}$$

Relative %Difference in Avg success rates

$$Rel\ \%Diff = \frac{\%S_T - \%S_C}{\%S_C}$$

THE MATH: *Statistical principles underlying Significance and Level of Confidence as it relates to online optimization testing*

Principle: Distinguish between Random Variation and a 'Real' systematic difference. I.e., what are the chances that what I am observing could happen even though the two treatments were really IDENTICAL?

Mathematical Method:

Determine:

What is the best measure of the **Dispersion (amt. of variation)** within the treatment performance observations?

Using: Standard Deviation (sd) of 'Difference in Sample Means'

$$sd = \sqrt{\frac{p_a(1-p_a)}{n_a} + \frac{p_b(1-p_b)}{n_b}}$$

and Presumption of Normal distribution of sample data

(this is not an unreasonable presumption for a well-designed optimization test, and allows us to use the Central Limit theorem)

Probability Density: Normal distribution

$$P(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

Margin of Error (2*sd)

$$d = 2\sqrt{\frac{p_a(1-p_a)}{n_a} + \frac{p_b(1-p_b)}{n_b}}$$

THE MATH: *Statistical principles underlying Significance and Level of Confidence as it relates to online optimization testing*

Principle: Distinguish between Random Variation and a 'Real' systematic difference. I.e., what are the chances that what I am observing could happen even though the two treatments were really IDENTICAL?

Mathematical Method:

Determine:

What is the likelihood that the observed difference could just be the result of random variation, even if the two treatments are identical?

Using: Statistical Level of Confidence

Confidence Interval

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

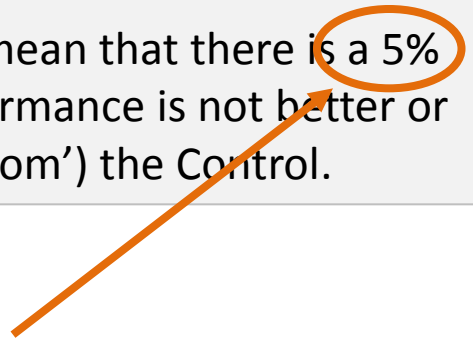
$\hat{p}$ is the proportion of successes
α is the error percentile
n is the sample size

Additional Notes:

- Though perhaps slightly counterintuitive, statistical confidence level is really aimed at answering the question: *"What are the chances that the apparent 'winning' Treatment is, in fact, **not** really better (…is perhaps even worse) than the Control"*

Example:

> A 95% level of confidence would mean that there is a 5% chance that the Treatment's performance is not better or worse than (i.e., is not 'different from') the Control.



- The complement of the statistical level of confidence is referred to as the **statistical significance level**. This is what is used in the MECLABS Test Protocol.

THE PURPOSE: *The purpose of Sample Distortion related validity calculations in Optimization is to enable you to …*

- ***Plan** Tests*
- ***Assess the outcome** of tests*
- ***Facilitate interpretation** of the outcome to improve you customer theory*

Objective: Distinguish between Random Variation and a 'Real' systematic difference.
I.e., what are the chances that what I am observing could happen even though the two treatments were really IDENTICAL?

Decide: What is the desired level of certainty?
How big a difference do I need to make it worthwhile to make a change?

Measure: Observed 'difference in sample means'
How many samples (trials, visits, …)?
How many 'successes' (conversions, clicks, …)?

Know: Your traffic rates (# visits per day, # in email list, …)
Number of treatments you intend to test

THE OBJECTIVE: *The purpose of Sample Distortion related validity calculations in Optimization is to enable you to : Plan, Assess, Facilitate interpretation …*

What you'll use the Validity math for:

Plan Tests:

How long should I expect the test to take if I achieve my desired min. difference amount?
What are my options if the estimated test duration is longer than I can afford to wait?

Assess Outcome of Tests:

Have I reached a sufficient sample size to achieve test validity?
Are the results of my test Conclusive?
(i.e., can I "Accept" or "Reject" my hypothesis based on the data I have collected to-date?)

Examples: Using MECLABS Test Protocol – Validity Tool

Key Principles

1. The four basic steps to establish sample size validity

   STEP 1: Estimate the success rate, sample size, and amount of difference

   STEP 2: Estimate the duration of the test

   STEP 3: Gather and record the actual data

   STEP 4: Determine if the data collected is sufficient to end the test

MECLABS Test Protocol – Validity Tool

Example Scenario

- You believe your primary offer page could convert better.
- You perform a diagnostic analysis on the existing offer page and arrive at a Hypothesis that form length and layout are contributing to excessive Friction.
- You design and develop **3 treatments** that will serve to test your hypothesis.
- Your recent **conversion rate** has been around **10%**.
- You want to be at least **95%** confident that the results of your test will **detect a difference** in conversion rate of **1% or more (10% relative)**.
- The average **traffic level** for this page has been about **500 visitors per day**.

How long will it take to achieve a valid sample?

Example Test Scenario

Minimum sample size estimation

Metric	Qty
Minimum size difference you wish to detect:	1.00%
Expected average success rate in the test:	10.0%
(Use 50% to be most conservative)	
Statistical significance level:	5%
(Remember, this is the probability of accidentally concluding there is a difference, due to sampling error, when really there is no difference.)	
Number of standard deviations for this significance level:	1.95996
Required standard deviation:	0.005102
Expected 2p(1-p)	0.18
Number of observations per treatment:	6,915

Test duration projection

Metric	Qty	
Arrival rate (arrivals/day)	500	arrivals/day
Number of treatments	4	
#Sample success observations / day	50	Successes/day
Estimated minimum acceptable sample size (#	27,659	Arrivals
Estimated minimum test duration (days)	55	days

Inputs

- **1.0%** is the minimum difference you wish to detect
- **10%** is your expected average conversion rate
- Since you decided on a 95% confidence level, you enter its complement, a **5%** statistical significance level
- Based on historical data, traffic level for the page is about **500** visitors per day
- Since you are testing three experimental treatment against the Control page, the total number of treatments is **4**

Output

- The estimated minimum test duration is calculated to be **55** days

Lets say you can only afford to wait 21 days. What are your options?

What are my options?

- Traffic? (X)
- # Experimental treatments? Try 2… Try 1.
- Confidence Level (1-sl)? Try 90%. (sl=10%)
- Bigger size difference? Try 20% instead of 10%.

Now how long is it expected to take?

Example Test Scenario (continued)

Test gets underway.

15 days later…

- Control: Visits: 3,728 Conversions: 371
- Treatment: Visits: 3,781 Conversions: 403

Enough data? Conclusive?

21 days later…

- Control: Visits: 5,971 Conversions: 587
- Treatment: Visits: 5,948 Conversions: 638

Enough data? Conclusive?

25 days later…

- Control: Visits: 6,488 Conversions: 657
- Treatment: Visits: 6,496 Conversions: 711

Enough data? Conclusive?

Validity for Decision Makers

Key Principles

Essential Definitions

Sample Distortion Effect – the effect on a test outcome caused by failing to collect a sufficient number of observations

Statistical Level of Confidence – the statistical probability that there really is a performance difference between the control and experimental treatments based upon the samples collected to date

Key Principles

1. There are essentially three factors that determine an outcome's statistical certainty:

   - **The Amount of Difference** – The magnitude of the difference in the primary metric between the Control and an Experimental treatment. (usually expressed as a percentage value)
   - **The Success Rate** – The number of experimental trials in which the outcome was favorable with respect to the primary success metric divided by the total samples received. (usually expressed as a percentage value)
   - **The Samples Received** – The total number of experimental trials used to compute the measured sample success rate. (e.g., number of visitors to a landing page, number of email messages sent).

2. An outcome's true statistical certainty is then compared with the desired level of certainty (chosen prior to experimentation).

Key Principles

1. The four basic steps to establish sample size validity

   STEP 1: Estimate the success rate, sample size, and amount of difference

   STEP 2: Estimate the duration of the test

   STEP 3: Gather and record the actual data

   STEP 4: Determine if the data collected is sufficient to end the test

Validity for Decision Makers

Thank You

Practical Example: MECLABS Test Protocol – Detail 1

RESEARCH • PUBLISHING • EDUCATION

ID: TP2047_Follow Up Radical Template

$u = 2q + t + m + 2v + i$ ©

1. Question

1.1) Primary Research Question	Which email design yields the highest click through rate?
1.2) Secondary Research Question	Which email design yields the highest conversion rate?

2. Treatments (These Variables / Values will determine the test treatments)

2.1) What is the variable?	Removal of top nav and improvements to layout, simplified approach
2.2) What are the values?	Control
	Treatment 1- w/2 side offers
	Treatment 2 - w/color background
	Treatment 3 - w/white background
	Treatment 4 - w/white background and side nav
	Treatment 5 - w/white background and colored nav
	Treatment 6- w/ white background & more products links

3. Metrics

3.1) What must we measure in order to determine the best performing value?

Metrics	Notes
Click-through Rate	
Conversion Rate	

4. Validation (Refer to separate Validation worksheet)

Note: Do not enter data in sections 4.2-4.4. They are linked to the separate Validation worksheet.

Practical Example: MECLABS Test Protocol – Detail 2

4. Validation (Refer to separate Validation worksheet)

Note: Do not enter data in sections 4.2-4.4. They are linked to the separate Validation worksheet.

4.1) Have you considered the impact of these validation threats?

Validity Threat Types	Considered (Y/N)	Describe Possible Threats For This Test
History effect	Yes	Product category is highly seasonal; especially with regard to holidays, et al. Normalizing using priors.
Instrumentation effect	Yes	3 text CTA's in each treatment share a common CTID -- Aggregating for CTR sub-analysis.
Selection effect	Yes	Using an asymmetrical send profile. Validation methodology is already profile-aware --> not a factor.
Sample Distortion effect	Yes	Splitter has been verified in prior tests using Control-Control method. Passed.
Other Validity effect	Yes	No known.

4.2) What is the sample size needed to ensure that the results will be predictive?

Projected minimum acceptable sample size:	38223	From validation worksheet - Primary Research Metric.

4.3) How long should it take to obtain the minimum sufficient sample size?

Estimated Test Duration	0	Days	From validation worksheet - Primary Research Metric.

4.4) How will you verify during the test that you have reached a sufficient number of sample observations?

Sample Size: (Validation worksheet)		**Notes**
Events	976517	
Successes	4955	
Sample Size Valid?:	NO	(based upon the corresponding Validation worksheet - Primary Research Metric.)

5. Metrics and Results

5.1) What are the test results?

Treatments	CTR Total	Relative Difference %	Validation Status	CR Total	Relative Difference %	Validation Status	Notes
Control	0.51%		YES	0.11%		YES	For the test cluster to-date, the Control is outperforming all treamtents at the aggregate level. A quick subsetting dive shows signif. Perf. Asymmetries. Now looking for corel-factors. (BK)
Treatment 1- w/2 side offers	0.44%	-14.22%	YES	0.09%	-17.96%	YES	
Treatment 2 - w/color background	0.42%	-16.80%	YES	0.09%	-15.60%	YES	
Treatment 3 - w/white background	0.46%	-10.55%	YES	0.11%	-3.81%	NO	
Treatment 4 - w/white background and side nav	0.50%	-1.31%	NO	0.10%	-7.14%	NO	
Treatment 5 - w/white background and colored nav	0.48%	-5.08%	NO	0.10%	-6.10%	NO	
Treatment 6- w/ white background & more	0.49%	-3.75%	NO	0.10%	-9.68%	NO	

6.Interpretation

Practical Example: MECLABS Test Protocol – Detail 3

5. Metrics and Results

5.1) What are the test results?

Treatments	CTR Total	Relative Difference %	Validation Status	CR Total	Relative Difference %	Validation Status	Notes
Control	0.51%		YES	0.11%		YES	For the test cluster to-date, the Control is outperforming all treamtents at the aggregate level. A quick subsetting dive shows signif. Perf. Asymmetries. Now looking for corel-factors. (BK)
Treatment 1- w/2 side offers	0.44%	-14.22%	YES	0.09%	-17.96%	YES	
Treatment 2 - w/color background	0.42%	-16.80%	YES	0.09%	-15.60%	YES	
Treatment 3 - w/white background	0.46%	-10.55%	YES	0.11%	-3.81%	NO	
Treatment 4 - w/white background and side nav	0.50%	-1.31%	NO	0.10%	-7.14%	NO	
Treatment 5 - w/white background and colored nav	0.48%	-5.08%	NO	0.10%	-6.10%	NO	
Treatment 6- w/ white background & more	0.49%	-3.75%	NO	0.10%	-9.68%	NO	

6.Interpretation

What insights can we gain from this test?

6.1) Objective Interpretation based on Test Data	For the test cluster to-date, the Control is outperforming all treamtents at the aggregate level. There is significant variance depending upon whether Product or Collection page is the landing page. Treatments 3 and 4 outperformed Control by 12% - 22% (+-3% MoE) in 3 of 4 sends when Category page is the landing page. [ref. pivot & correlation tabs]. (BK)
6.2) Expert Speculation on Possible causes of the Test Outcome	A quick subsetting dive shows significant performance asymmetries, with periods during which treatments (esp. Treatment 3) substantially outperforms Control. Now looking for potential corellative factors. First set includes whether landing page is Product or Collection page (link study). (BK)
6.3) Suggested Follow-Up Tests	