W241 Final Project: eBags Feature Vision Experiment

Anusha Munjuluri, Mike Frazzini, and Raymond Lee

4/26/2018

Abstract

In E-commerce, products only exist in a digital representation until they are purchased and delivered. Product information, such as features and benefits, is important to online consumers and drives purchase behavior. This useful information that exists today in textual form, can be presented in a richer, integrated fashion within product images and media. This drives more customers to see and understand product benefits, and potentially add more products to their online carts and by extension, purchase more products. Our experimental design takes 4 widely visited and popular bag models from a well-known E-commerce site, eBags, and implements a treatment of integrated product features directly within product imagery of the models' product detail pages. Our experiment revealed that there are promising opportunities for improvement in the display of product information in imagery that may have a meaningful causal effect on orders and revenue, as well as driving improvements in how eBags does experiments with A/B split testing.

1 Introduction

E-commerce has indelibly changed retail in many ways, and one of the biggest ways is by stretching out the tail of product availability and setting the bar high for product information and expertise. Chris Anderson, former Wired editor and technology pundit, expertly describes this phenomenon in his book, The Long Tail: Why the Future is Selling Less of More [1]. One of the challenges the long tail has created for retailers is collecting and extending product information, providing easy to use information and search tools for making sense of all the product information, among vast arrays of product assortments. eBags, Inc. is a successful specialty e-retailer featuring the world's top brands and over 90,000 products in luggage, bag, and travel accessories categories. eBags has amassed extensive product information for all of its products to help people find the perfect bag and accessory for their journeys and adventures. As it continues to strive to be the best e-retailer in its categories, eBags is asking the question, "is there a better way to present product information to help our customers find their perfect bags and accessories?"

[1] Anderson, "The Long Tail"

2 Experiment Design

2.1 Model Selection:

Four top bag models were chosen based on high visit levels (site traffic). This ensures that there would be enough page visits for each model to provide for a successful experiment as well as to achieve a high statistical power level (see Statistical Power section for more on this) for our results. The team also felt it was important to choose bags from several categories such as luggage, backpacks, travel handbags and not restrict our experiment to products from just one category. This also allowed us to increase generalizability of our experiment for several types of products and potentially different types of customers that are interested in different types of bags.

The four products selected are listed below. They can be seen in the image below (Figure 1):

- eBags TLS Motherlode Mini
- Travelon Wheeled Underseat Tote
- Piazza Cross-Body Handbag
- eBags Professional Weekender Backpack

Figure 1: 4 models selected: TLS, Wheeled Travelon, Crossbody Piazza, Professional Weekender

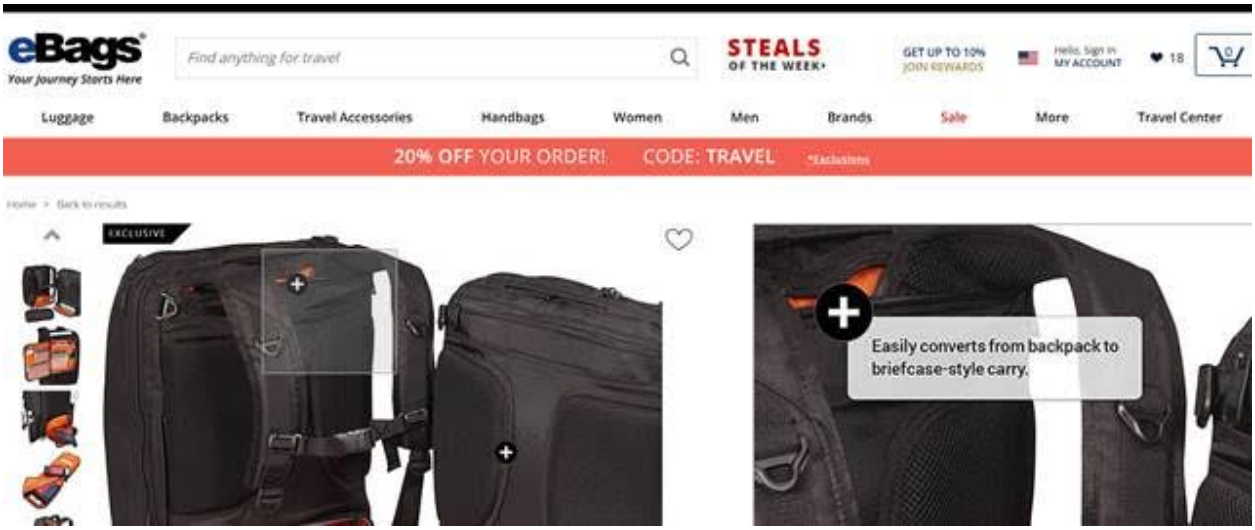

Figure 2: Treatment with integrated product features in product images

In the experiment, a visitor to the eBags Product Detail Page (PDP) for any of the 4 models selected, will either be shown the control of the existing page containing images of a bag with product features and benefits listed in textual format much lower on the page, or the intervention of a new page that has the most relevant and important product features integrated and highlighted in context within product images. Image shown above presents an example of how the treatment looks (Figure 2).

In control, product features and benefits text is much lower on the page as per the status quo and is not integrated into the product imagery like the treatment. Current control state, without the feature/benefit callouts and text integrated into the main imagery, can be seen here:

https://www.ebags.com/product/samsonite/spinner-underseater-with-usb-port-ebags-exclusive/334846?productid=10525770

Image below presents an example of how control looks (Figure 3).

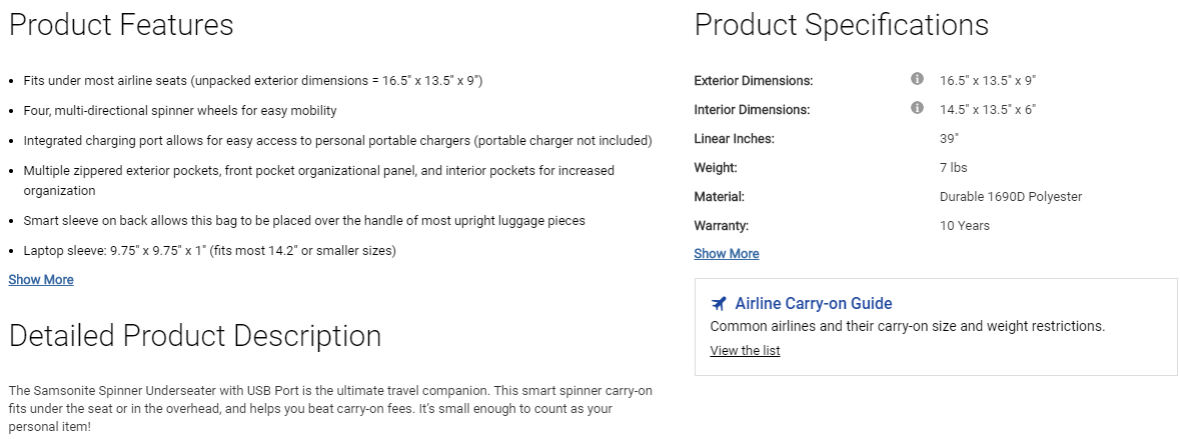

Figure 3: Control features shown at bottom of product detail page (PDP)

2.2 Hypothesis:

Ho: Null Hypothesis: Add-to-Cart rate (ATC) is same for both treatment and control for the 4 products in test. Baseline ATC is ~16% for the 4 models.

Ha: Alternate Hypothesis: ATC will be >=5% more (of baseline) for treatment intervention of feature callouts within product detail images of the 4 bag models than control.

eBags team considers a 5% increase in the ATC or the orders to be a significant effect of interest and a promising opportunity for improving feature vision of products. 5% of current baseline for ATC puts ATC in a range of (15.2% - 16.8%). Any effect larger than 0.8% in either direction, would show us that the treatment had a minimum detectable effect of 5% in the experiment, over the current baseline.

We are also interested in the outcome of order rates (conversion/CV) between treatment and control which has a baseline of 6% (Range: 5.7% - 6.3%) for PC devices and 3% (Range: 2.85% - 3.15%) for mobile devices. Ranges for these baseline rates have been calculated using a power calculator (see Statistical Power section for more on this).

2.3 Outcome Measures:

Outcome measures of interest for this experiment are: **ATC** (Add-to-Cart) and **Orders** of the 4 models whose PDP were changed. We were interested in measuring if the intervention of integrated product features caused an increase in the current ATC or orders rate for the treatment group. ATC and Orders are numeric measures (that is, 1 if they ordered a product, 0 if they did not order). Values can be greater than 1 if more than one product was ordered or added-to-cart in a visit.

2.4 Covariates of Interest:

Covariates of interest are:

- Device (mobile, tablet, pc)
- New or recent user
- Marketing channels (primarily: Email, Keywords, SEO, Social, Untracked).

Device: Integrating product features in product images creates a new user experience for the users. We were interested in seeing if this change had different (heterogeneous) effects on different devices. For example:

is it easier to zoom or hover over product images to see the integrated features using PC or mobile phones? Is the usability and user experience consistent or significantly different across devices?

New vs Recent Users: When a user visits eBags for the first time, they are considered as a new user and a new cookie is placed on their device for eBags site. When a new user visits the site for the second time or existing users/customers visit eBags site, they are called recent users. We wanted to measure if the treatment made any difference in ATC and orders for unaccustomed users compared to recent users.

Marketing Channels: There are 11 marketing channels through which eBags get user traffic such as emails, advertising, affiliates etc. Of those 11 channels, primary channels of interest and those which generate the most traffic are: Email, Keywords, SEO (Search Engine Optimization), Social and Untracked (logging onto eBags site from browser directly). Conditional on marketing channel, we wanted to see if there was any significant difference in the treatment effect between these various channels.

2.5 Blocking and Clustering:

We specifically didnt have to block for any of the covariates of interest because number of visits to the 4 models selected is very high about ~86K records per week, for the four models out together. When sample size is high and we are splitting into control and treatment on visiting a model PDP, we werent worried about a specific covariate being more assigned to treatment or control (i.e more correlated with treatment). Our covariate balance check (See Covariate Balance section) confirms this belief of ours.

- ~ 29,000 PDP Visitors Per Week for Professional Weekender Bag
- ~ 22,000 PDP Visitors Per Week for Wheeled Underseat Bag
- ~ 18,000 PDP Visitors Per Week for TLS Bag
- ~ 17,500 PDP Visitors Per Week for Piazza Bag

We didn't have any clustering effects to consider because this web test was not targeted towards a specific group of people or locations but was administered to anyone who visited any of the 4 models on eBags site.

2.6 Experiment Type and ROXO Grammar

This is a between subjects experiment as we are comparing ATC and order rates between treatment and control groups. We have an RXO experiment design where we first randomize, show changed PDP to treatment group and then observe the outcomes in treatment and control groups.

- Control: **R**andomize – **O**bserve
- Treatment: **R**andomize **X**periment **O**bserve

2.7 Statistical Power (Pre-Treatment):

Before running the experiment, we calculated sample size needed for each outcome measure: ATC rate and order rate to ensure we have enough statistical power in our results. Note: We used the statistical power calculator from http://www.evanmiller.org/ab-testing/sample-size.html for our calculations.

Statistical power calculation for ATC (shown in Figure 4):

- The baseline ATC conversion rate average for these models, as tracked historically by eBags, is 16%.
- We determined our minimum detectable effect to be 5%.
- We determined our desired statistical significance level as 5%.
- We determined our desired statistical power as 80%.
- We estimated our total sample size to be ~66k. (33k per variation)

Statistical power calculation for Orders (shown in Figure 5):

- The baseline order conversion rate, as tracked by eBags, is 6%.

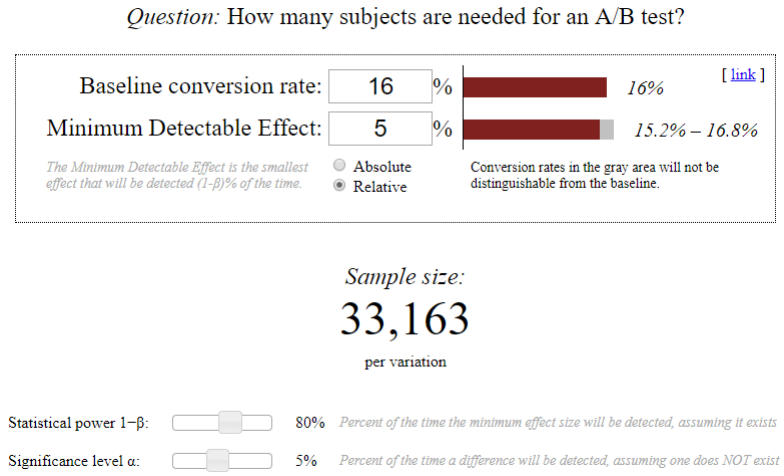

Figure 4: Power Calculation for ATC: Pre-Treatment

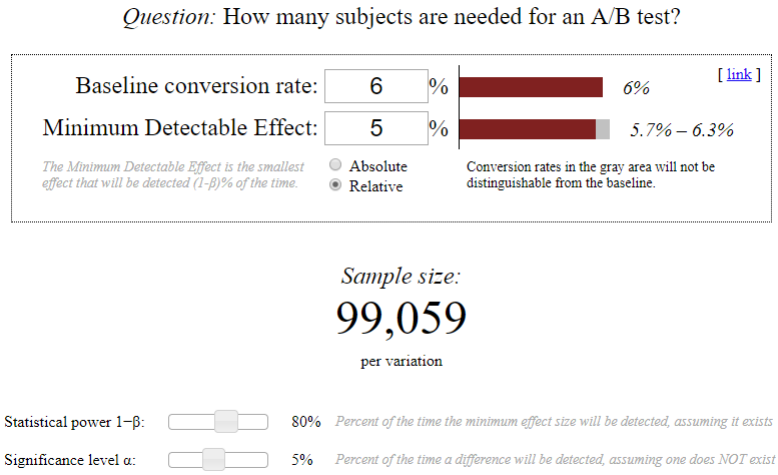

Figure 5: Power Calculation for Order: Pre-Treatment

- We determined our minimum detectable effect to be 5%.
- We determined our desired statistical significance level as 5%.
- We determined our desired statistical power as 80%.
- We estimated our total sample size to be ~200k. (99k per variation)

3 Experiment Web Test Implementation:

3.1 Determining Models' Feature Vision:

For the four products included in the test, in order to achieve a treatment intervention that would have a high potential for success, it was important to capture the most important product features for each model. Several hours were spent by the team reviewing product feature text, product feature videos, and also leveraging basic Natural Language Processing (NLP) techniques that combed through ~6,000 customer product reviews for the four models. This helped us to identify the top features and benefits that customers were calling out. Mock ups were created for all four products that showed highlight placement and feature text on each of the 5-7 product images per model.

NLP processing techniques used were as follows:

Python script and procedures written to parse and process text of over 6,000 customer product reviews leveraging the python nltk library:

- For each string of text:

  i.) Remove all punctuation

  ii.) Remove all stop-words

  iii.) Return cleaned text as a list of words
- Utilize a word vectorizer and sparse matrix to create ngrams of 3-6 word phrases
- Build a frequency table and sort to display the most common word phrases
- Review manually to pull top features/benefits text

3.2 Randomization

Randomization approach for this experiment was to randomly assign a visitor to control or treatment, the first time they visit any of the 4 product detail pages. This was done by a random number generator in the code that dynamically displays the control page (unchanged) or the treatment page with the intervention to subjects in the experiment. From that point on, the session and visitor behavior was logged in association with the group (treatment/control) they are assigned to along with their unique cookie id. This method, like most website tracking, is reliant on the visitor enabling and accepting web cookies. Using cookies is the most common mechanism used for tracking session and session state on the web because of stateless http protocol[2]. This aligns with common and generally prescribed approaches to controlled experiments on the web[3] which are also referred to as "split-tests" or "A/B tests."

Irrespective of the 4 models a user clicks on for the first time, once assigned to treatment, they will be in treatment for the rest of the models as well. Cookies ensure that assignment to treatment or control persists over time. That is, every user assigned to treatment is always in treatment unless they delete their cookies or switch devices.

It should be noted that this is a "Server-side" split test experiment that utilizes a random number generator in the web server application to assign each visitor based on a 25/25/50 split (control A1/control A2/treatment B1). Control or Treatment is *randomly assigned* on first PDP visit based on a random number generator reflecting a 25/25/50 split.

Some additional notes on A/B testing:

- Assignment persists via cookie variable.
- Assignment not preserved across devices (some spillover may occur - for more see Limitation of Device Attribution section.)

Client-side vs Server-side strengths and limitations:

- Client-side: Generally easier to implement but can cause "flicker" and slowness and not appropriate for complex functional tests (like a shopping cart test).
- Server-side: Generally better for performance and complex functional tests but harder to implement and requires a platform capability for server-side split-testing

[2] RFC 7230, "Hypertext Transfer Protocol."
[3] Kohavi, et. al.,

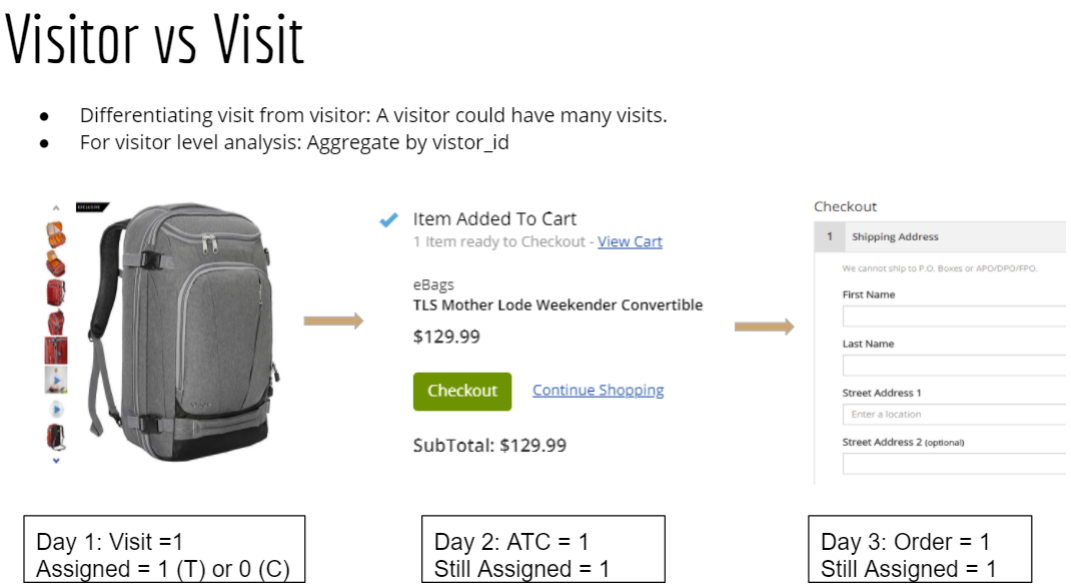

Figure 6: Visitor vs Visit

3.3 Why run a Dual-Control (Placebo Design)?:

A dual-control, aka placebo design (25/25/50 split), was used as part of eBags standard practice to allow for basic randomization validation and to provide an easy reference for the level of variance across tests (i.e. there should be very little difference between both control groups.) This also serves as a rudimentary power control as tests typically run until there is reasonably low variance across both control groups.

3.4 Understanding Visitor vs Visit:

Generally, the shopping experience for retail users consists of three phases: Viewing and comparing products, adding-to-cart products they are interested in and eventually ordering one or more of the products . This process however could spread over several days and may not finish in a short matter of time. We have noticed a similar behavior of users visiting eBags site as well. It is possible for a visitor to visit the site on multiple days i.e. they can view the product detail page one day, add-to-cart another day and order on another day.

eBags tracks users' visits and collects data about each visit, along with cookies assigned to each user. If a visitor visits the site on multiple days, there will be multiple visit records collected for the same user, tracking their activity on each day. Visitor's activity such as which products were viewed, added-to-cart or orders made are tracked. Common link joining all these multiple visits is the unique cookie id that each user gets when they visit eBags site for the first time. Cookies ensure that once a visitor is assigned to treatment, they remain in treatment till they delete their cookies (Figure 6). Treatment is preserved over multiple visits/days with the help of cookies. By grouping visit level records using cookie id, a wholistic view of a single visitor's shopping experience on eBags site can be obtained.

3.5 Pilot Testing

Pilot testing was done in the form of usability (UX) walkthroughs with several designers and one UX architect on staff at eBags. This "real-world" experiment involved and required collaboration from several different groups including Merchants, Web Designers, Web Developers, and other Information Technology (IT) personnel. Several different designs were proposed and mock-ups were created. Mock-ups were used to

do walkthroughs to help achieve the right balance of UX, time and complexity to implement. Some of the most significant improvements that came out of the walkthroughs are:

- Feature/benefit text was too cluttering to show on the initial main images and it was decided that it should only be shown on zoom/tap-to-zoom on mobile.
- Feature/benefit call-out icons needed to be on the initial main images (in addition to zoom), so that the visitor would be aware that there is more information about the product available within the images.
- Various icons for the call-outs were reviewed before the final ones were chosen.
- Various features/benefit text and call-out placements were changed.

Mock-up Microsoft PowerPoint slides can be reviewed here:

https://drive.google.com/drive/folders/13kTfClRzmZYB36T0OItPmkw9tS8bMagy?usp=sharing

3.6 Administering the experiment:

After inital pilot test, these 4 models' PDPs were changed according to the UX team guidelines and implemented by eBags feature development team. These changes were implemented in the following phases and their timeline is shown below:

- Feature Research: Week of March 11,2018
- Feature Development: Week of March 18,2018
- Experiment Pilot: Week of March 25,2018
- Experiment Go-Live: Week of April 1,2018
- Data Collected: April 6 - April 23,2018

Keeping the statistical power calculation in mind, we ran the experiment long enough to get a sample size which gives at least 80% statistical power for ATC.

4 Data Analysis

4.1 Data Collection:

Experiment was run for 18 days (April 6 - April 23, 2018) and 150,508 (150k) rows of data was collected with 44 variables of interest.

Important Variables

For each record (a.k.a visit), we got the following variables:

- *visitor_id* (cookie id),
- *request_dt* (date of visit)
- *partition* (i.e. split a1, split a2, b2 treatment)
- *visit_tls*, *visit_wheeled*, *visit_prof*, *visit_piazza* (if the user visited any of the 4 models or not, encoded as 0/1)
- *atc_tls*, *atc_wheeled*, *atc_prof*, *atc_piazza* (if the user added-to-cart any of the 4 models or not, encoded as 0/1)
- *ord_tls*, *ord_wheeled*, *ord_prof*, *ord_piazza* (if the user ordered any of the 4 models or not, encoded as 0/1)
- *device* (mobile, pc, tablet)
- *marketing_channel* (11 categories)
- *new_user* (1 if new)
- Other variables such as: *browser*, *operating_system*, *units_purchased* etc.

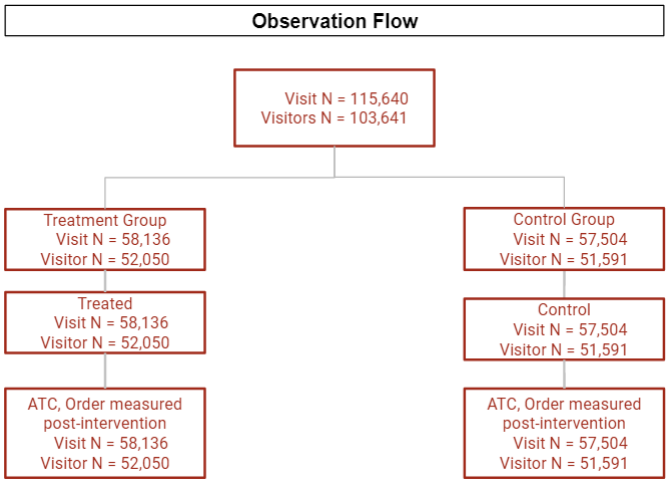

Figure 7: Observation Flow

4.2 Observation Flow

Figure 7 shown above is a visual of the experiment observation flow. Our experiment followed a "post-test control group" (RXO: Randomize, Xperiment, Observe) experiment design:

- Visitors would be randomly assigned to control or treatment on their first visit to one of the four experiment product detail pages
  - 50% of visitors were assigned to control, and 50% of visitors were assigned to treatment.
- The control group would then be exposed to the normal product detail page, and the treatment group would then be exposed to a feature vision product detail page
- At the end of our experiment period, we observed ATC and order rates for these two groups.

4.3 Covariate Balance Check

We assigned product detail page visitors to either Control: A1, Control: A2, or Treatment: B1 using a respective 25%/ 25%/50% split. We checked our data for covariate balance. As the table shows below, most covariates were a 25%/25%/50% split, with a few covariate groups off by 1 percentage point. This confirms that our randomization worked properly and there is no bias towards any covariates. A dual-control allowed basic randomization validation and to provide an easy reference for the level of variance accross tests.

Covariate Balance for each of the covariates has been shown in Figure 8 for the three groups: Control: A1, Control: A2, or Treatment: B1.

Covariates	Control: A1	Control: A2	Treatment : B1
Devices - Mobile	25%	25%	50%
Devices - PC	25%	25%	51%
Devices - Tablet	24%	25%	51%
Marketing - Social	24%	25%	51%
Marketing - Keyword	24%	26%	51%
Marketing - Email	25%	25%	50%
User - New	25%	25%	50%
User - Repeat	25%	25%	50%
Model - Wheeled	25%	25%	50%
Model - TLS	25%	25%	50%
Model - Professional	24%	25%	51%
Model - Piazza	25%	25%	50%

Figure 8: Covariate Balance

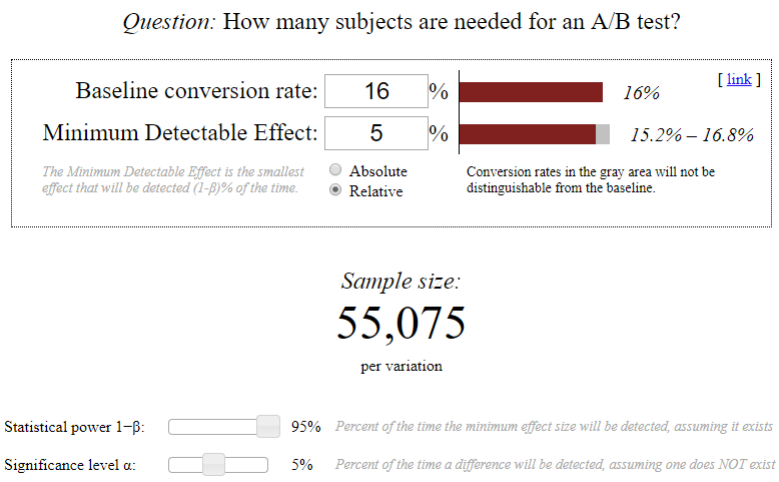

Figure 9: Power Calculation for ATC: Post-Treatment

4.4 Statistical Power (Post-Treatment):

- Statistical power for ATC outcome is ~95%% post treatment (shown in Figure 9):

  ~66k sample size was needed for ATC to detect a minimum effect of 5%, over the baseline of 16%, with 80% statistical power. Post treatment we had ~115k records. This gives us ~95% statistical power for ATC.

- Statistical power of order rate is ~60% post treatment (shown in Figure 10):

  ~200k samples size was needed for orders to detect a minimum effect of 5 % over the baseline of 6% with 80% statistical power for PC devices. Post treatment we had ~115k records. This gives us ~60% statistical power for orders.

  Note: We used the statistical power calculator from http://www.evanmiller.org/ab-testing/sample-size.html for our calculations.

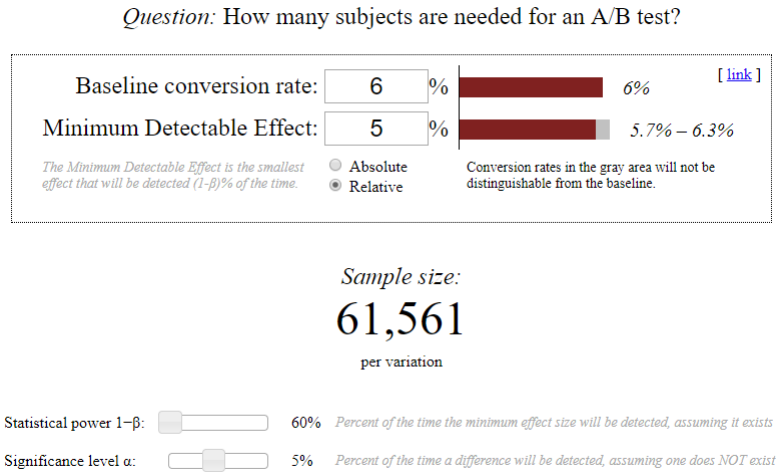

Figure 10: Power Calculation for Order: Post-Treatment

4.5 Data Preparation for Analysis:

For doing regression analysis, we took the following steps to prepare our data:

Variable Transformations:

- Encoded factor variables as dummy variables ex. *device* as *device_pc*, *device_tablet*, *device_mobile*
- Converted *partition* (treatment assignment) to a dummy variable: *assigned* i.e. *assigned* = 1 if in Treatment, = 0 if in Control
- For each record (visit), aggregated visits, atc and orders of all the 4 models as aggregate variables (*visit_aggregate*, *atc_aggregate*,*ord_aggregate*). These variables tell us how many of the 4 models a sinle user visited or added-to-cart or ordered in a single visit.
- *atc_aggregate* and *ord_aggregate* are the outcome variables used for regression.

Visitor Data: Aggregation by visitor

We got data at visit level from eBags team. To get visitor level data, we grouped visits by *visitor_id* and aggregated outcome measures of multiples visits for a single user. After aggregation by visitor id, we had ~105K records.

- Numerical columns such as visits, ATC and orders were summed across multiple visits for a single user
- **First-click Device Attribution:** We took the first click attributes for covariates that can change on multiple visits ex. Marketing Channels, New Users. That is, if they visited eBags site for the first time via email but subsequently using the browser or other channels, we considered their first visit channel of 'email' as the main attribute for that visitor for further analysis.

Removing subjects not in the experiment: Treated or not?

In eBags site, it is possible to directly add a product to cart from the home page, without viewing the product detail page as shown below (Figure 11). We wanted to include only those visitors who have actually seen the treatment by clicking on any of the 4 models PDP and are therefore, part of the experiment. At visit level we eliminated those records from the analysis, where the users did not visit nor add-to-cart nor order any of the models. At visitor level, we eliminated those records of users, who did not visit any of the 4 models because if they did not visit any of the 4 models, it means they never actually got assigned to treatment or control. Hence, such records are not part of the experiment. After elimination of such records, we were left with ~103k records at vistor level and ~115k records at visit level.

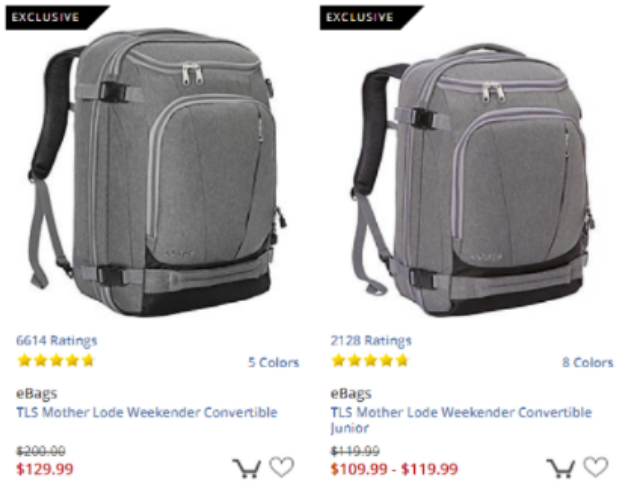

Figure 11: Adding products to cart from Home page

4.6 Outcomes Distribution:

Outcomes distribution at visit level is as follows:

- ATC: 2063 (Control), 2126 (Treatment)
- Orders: 1251 (Control), 1389 (Treatment)

5 Treatment Effect Analysis:

5.1 Regression Models:

Outcome measures: ATC and Orders were regressed on assigned variable, covariates of interest and their interaction terms. Regression equations for ATC are shown below. We split our analysis into various models because there are three different kinds of covariates of interest (devices, new users and marketing channels). We wanted to analyze each covariate scenario separately and avoid confounding multiple interaction terms at once.

Similar to ATC, orders have been regressed using below equations. These regressions were done at both visit and visitor level.

- **Model 1: Just assigned variable!**

  $ATC = \beta_0 + \beta_1 assigned$

- **Model 2: Device covariates and their interaction terms**

  Baseline Category: Device PC
  $ATC = \beta_0 + \beta_1 assigned + \beta_2 device\_tablet + \beta_3 device\_mobile + \beta_4 device\_tablet * assigned + \beta_5 device\_mobile * assigned$

- **Model 3: New user covariates and their interaction terms**

  Baseline Category: Recent users
  $ATC = \beta_0 + \beta_1 assigned + \beta_2 new\_user + \beta_3 new\_user * assigned$

- **Model 4: Marketing channels and their interaction terms**

  Baseline Category: Keywords

Other important channel categories included: Email, SEO, Social, Untracked

All remaining secondary channels combined under: mkt_others variable
$ATC = \beta_0 + \beta_1 assigned + \gamma \sum_{5} mkt\_channel + \alpha \sum_{5} assigned * covariate$

- **Model 5: Throwing the kitchen sink in! All covariates and their interaction terms**

  $ATC = \beta_0 + \beta_1 assigned + \beta_2 new\_user + \beta_3 device\_tablet + \beta_4 device\_mobile + \gamma \sum_{5} mkt\_channel + \alpha \sum_{8} assigned * covariate$

Below regression tables show outcomes of each of these 5 models for ATC and orders, at visit and visitor level. In the tables, column (1) refers to Model 1, columns (2), (3) to Model 2, columns (4), (5) to Model 3, columns (6), (7) to Model 4, columns (8), (9) to Model 5. For each of the models, first column shows regression equations with just covariates added and the second column with covariates and their interaction terms added.

5.2 Regression Tables:

Table 1: Regression of ATC at Visit level

	Dependent variable:								
	atc_aggregate								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
assigned	0.001	0.001	0.0003	0.001	0.001	0.0003	0.005	0.0003	0.006
	(0.001)	(0.001)	(0.003)	(0.001)	(0.002)	(0.001)	(0.003)	(0.001)	(0.007)
device_mobile		−0.053***	−0.053***					−0.031***	−0.032***
		(0.002)	(0.002)					(0.002)	(0.002)
device_tablet		−0.038***	−0.038***					−0.029***	−0.029***
		(0.002)	(0.003)					(0.002)	(0.003)
assigned:device_mobile			0.0004						0.0004
			(0.003)						(0.003)
assigned:device_tablet			−0.00005						−0.0001
			(0.004)						(0.004)
new_user				−0.025***	−0.025***			−0.020***	−0.019***
				(0.001)	(0.002)			(0.001)	(0.002)
assigned:new_user					−0.001				−0.002
					(0.002)				(0.002)
mkt_email						0.0001	0.004	−0.008*	−0.003
						(0.004)		(0.004)	(0.006)
mkt_seo						0.014**	0.018	0.007	0.011
						(0.005)		(0.005)	(0.007)
mkt_social						−0.067***	−0.065	−0.059***	−0.057***
						(0.003)		(0.003)	(0.005)
mkt_untracked						0.014**	0.015	0.013**	0.014*
						(0.004)		(0.004)	(0.006)
mkt_others						−0.045***	−0.044	−0.051***	−0.050***
						(0.004)		(0.004)	(0.005)
assigned:mkt_email							−0.009		−0.009
									(0.008)
assigned:mkt_seo							−0.007		−0.008
									(0.010)
assigned:mkt_social							−0.005		−0.005
									(0.007)
assigned:mkt_untracked							−0.002		−0.002
									(0.008)
assigned:mkt_others							−0.003		−0.003
									(0.007)
Constant	0.036***	0.072***	0.072***	0.050***	0.049***	0.076***	0.074***	0.107***	0.104***
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.003)	(0.002)	(0.004)	(0.005)
Observations	115,640	115,640	115,640	115,640	115,640	115,640	115,640	115,640	115,640
R^2	0.00000	0.015	0.015	0.004	0.004	0.032	0.032	0.039	0.040
Adjusted R^2	−0.00001	0.015	0.015	0.004	0.004	0.032	0.032	0.039	0.039
Residual Std. Error	0.189 (df = 115638)	0.188 (df = 115636)	0.188 (df = 115634)	0.189 (df = 115637)	0.189 (df = 115636)	0.186 (df = 115633)	0.186 (df = 115628)	0.186 (df = 115630)	0.186 (df = 115622)

Note: *p<0.05; **p<0.01; ***p<0.001

Table 2: Regression of Orders at Visit level

	Dependent variable:								
	ord_aggregate								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
assigned	0.002*	0.002*	0.005*	0.002*	0.004*	0.002*	0.010*	0.002*	0.014**
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.005)	(0.001)	(0.005)
device_mobile		−0.028***	−0.026***					−0.021***	−0.020***
		(0.001)	(0.002)					(0.001)	(0.002)
device_tablet		−0.021***	−0.020***					−0.020***	−0.018***
		(0.002)	(0.002)					(0.002)	(0.002)
assigned:device_mobile			−0.004						−0.003
			(0.002)						(0.002)
assigned:device_tablet			−0.003						−0.003
			(0.003)						(0.003)
new_user				−0.021***	−0.019***			−0.021***	−0.020***
				(0.001)	(0.001)			(0.001)	(0.001)
assigned:new_user					−0.003				−0.003
					(0.002)				(0.002)
mkt_email						−0.007*	−0.003	−0.015***	−0.010**
						(0.003)	(0.004)	(0.003)	(0.004)
mkt_seo						0.004	0.008	−0.001	0.003
						(0.004)	(0.005)	(0.004)	(0.005)
mkt_social						−0.026***	−0.022***	−0.021***	−0.017***
						(0.002)	(0.003)	(0.002)	(0.003)
mkt_untracked						0.009**	0.012**	0.009**	0.012**
						(0.003)	(0.004)	(0.003)	(0.004)
mkt_others						−0.022***	−0.018***	−0.029***	−0.024***
						(0.003)	(0.004)	(0.003)	(0.004)
assigned:mkt_email							−0.008		−0.009
							(0.006)		(0.006)
assigned:mkt_seo							−0.007		−0.008
							(0.007)		(0.007)
assigned:mkt_social							−0.009		−0.009
							(0.005)		(0.005)
assigned:mkt_untracked							−0.007		−0.007
							(0.006)		(0.006)
assigned:mkt_others							−0.009		−0.009
							(0.005)		(0.005)
Constant	0.022***	0.041***	0.040***	0.033***	0.032***	0.039***	0.035***	0.064***	0.058***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.003)	(0.003)	(0.004)
Observations	115,640	115,640	115,640	115,640	115,640	115,640	115,640	115,640	115,640
R^2	0.00005	0.007	0.007	0.005	0.005	0.008	0.008	0.016	0.016
Adjusted R^2	0.00004	0.007	0.007	0.005	0.005	0.008	0.008	0.016	0.016
Residual Std. Error	0.151 (df = 115638)	0.150 (df = 115636)	0.150 (df = 115634)	0.150 (df = 115637)	0.150 (df = 115636)	0.150 (df = 115633)	0.150 (df = 115628)	0.150 (df = 115630)	0.150 (df = 115622)

Note: *p<0.05; **p<0.01; ***p<0.001

Table 3: Regression of ATC at Visitor level

	Dependent variable:								
	atc_aggregate								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
assigned	0.001	0.001	0.0003	0.001	0.003	0.0004	0.009*	0.0005	0.011
	(0.001)	(0.001)	(0.004)	(0.001)	(0.003)	(0.001)	(0.004)	(0.001)	(0.009)
device_mobile		−0.060***	−0.060***					−0.035***	−0.035***
		(0.002)	(0.003)					(0.002)	(0.003)
device_tablet		−0.043***	−0.044***					−0.033***	−0.033***
		(0.003)	(0.004)					(0.003)	(0.004)
assigned:device_mobile			0.0005						0.0002
			(0.004)						(0.004)
assigned:device_tablet			0.0003						−0.001
			(0.005)						(0.005)
new_user				−0.030***	−0.028***			−0.025***	−0.023***
				(0.001)	(0.002)			(0.002)	(0.002)
assigned:new_user					−0.003				−0.004
					(0.003)				(0.003)
mkt_email						0.006	0.013	−0.004	0.004
						(0.005)		(0.005)	(0.007)
mkt_seo						0.020**	0.028	0.012	0.020*
						(0.006)		(0.006)	(0.009)
mkt_social						−0.073***	−0.069	−0.065***	−0.061***
						(0.004)		(0.004)	(0.005)
mkt_untracked						0.019***	0.022	0.018***	0.021**
						(0.005)		(0.005)	(0.007)
mkt_others						−0.049***	−0.046	−0.057***	−0.054***
						(0.004)		(0.004)	(0.006)
assigned:mkt_email							−0.014		−0.016
									(0.010)
assigned:mkt_seo							−0.015		−0.016
									(0.012)
assigned:mkt_social							−0.008		−0.009
									(0.008)
assigned:mkt_untracked							−0.006		−0.006
									(0.010)
assigned:mkt_others							−0.006		−0.007
									(0.009)
Constant	0.040***	0.081***	0.081***	0.058***	0.057***	0.083***	0.079***	0.121***	0.116***
	(0.001)	(0.002)	(0.003)	(0.001)	(0.002)	(0.004)	(0.003)	(0.004)	(0.006)
Observations	103,641	103,641	103,641	103,641	103,641	103,641	103,641	103,641	103,641
R^2	0.00000	0.014	0.014	0.004	0.004	0.031	0.031	0.039	0.039
Adjusted R^2	−0.00001	0.014	0.014	0.004	0.004	0.031	0.031	0.038	0.038
Residual Std. Error	0.221 (df = 103639)	0.219 (df = 103637)	0.219 (df = 103635)	0.220 (df = 103638)	0.220 (df = 103637)	0.217 (df = 103634)	0.217 (df = 103629)	0.216 (df = 103631)	0.216 (df = 103623)

Note: *p<0.05; **p<0.01; ***p<0.001

Table 4: Regression of Orders at Visitor level

	Dependent variable:								
	ord_aggregate								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
assigned	0.003*	0.002*	0.006*	0.003**	0.005**	0.002*	0.011*	0.002*	0.018**
	(0.001)	(0.001)	(0.003)	(0.001)	(0.002)	(0.001)	(0.005)	(0.001)	(0.006)
device_mobile		−0.032***	−0.030***					−0.024***	−0.022***
		(0.001)	(0.002)					(0.001)	(0.002)
device_tablet		−0.024***	−0.023***					−0.022***	−0.021***
		(0.002)	(0.003)					(0.002)	(0.003)
assigned:device_mobile			−0.005						−0.004
			(0.003)						(0.003)
assigned:device_tablet			−0.003						−0.003
			(0.004)						(0.004)
new_user				−0.022***	−0.019***			−0.023***	−0.020***
				(0.001)	(0.002)			(0.001)	(0.002)
assigned:new_user					−0.005*				−0.005*
					(0.002)				(0.002)
mkt_email						−0.006	−0.001	−0.014***	−0.008
						(0.003)	(0.004)	(0.003)	(0.004)
mkt_seo						0.007	0.011	0.001	0.006
						(0.004)	(0.006)	(0.004)	(0.006)
mkt_social						−0.029***	−0.024***	−0.023***	−0.019***
						(0.003)	(0.004)	(0.003)	(0.004)
mkt_untracked						0.013***	0.017***	0.013***	0.016***
						(0.004)	(0.005)	(0.004)	(0.005)
mkt_others						−0.026***	−0.020***	−0.033***	−0.027***
						(0.003)	(0.004)	(0.003)	(0.004)
assigned:mkt_email							−0.010		−0.012
							(0.006)		(0.006)
assigned:mkt_seo							−0.008		−0.009
							(0.008)		(0.008)
assigned:mkt_social							−0.010		−0.009
							(0.005)		(0.005)
assigned:mkt_untracked							−0.007		−0.008
							(0.007)		(0.007)
assigned:mkt_others							−0.011		−0.012*
							(0.006)		(0.006)
Constant	0.024***	0.046***	0.045***	0.037***	0.036***	0.042***	0.038***	0.072***	0.064***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.003)	(0.004)	(0.003)	(0.004)
Observations	103,641	103,641	103,641	103,641	103,641	103,641	103,641	103,641	103,641
R^2	0.0001	0.008	0.008	0.004	0.005	0.009	0.010	0.018	0.018
Adjusted R^2	0.0001	0.008	0.008	0.004	0.005	0.009	0.009	0.018	0.018
Residual Std. Error	0.161 (df = 103639)	0.161 (df = 103637)	0.161 (df = 103635)	0.161 (df = 103638)	0.161 (df = 103637)	0.161 (df = 103634)	0.161 (df = 103629)	0.160 (df = 103631)	0.160 (df = 103623)
Note:									*p<0.05; **p<0.01; ***p<0.001

5.3 Models Analysis (Findings):

- **Covariate Balance Check:**

  For ATC and Orders, ATE stayed at 0.001 and 0.002 respectively, even after adding covariates (Refer Regression Tables 1,2). This gave us confidence that our randomization worked and there is no imbalance in our covariates. ATC coefficient for *assigned* variable slightly shifts when marketing channel covariates are added. This is probably because of the large number of marketing channel categories available (11) and because it is hard to control for a perfect 25/25/50 split in some of the minor channels. All the major channels have covariate balance as shown in the covariate balance section.

- **ATC ATE:**

  Our alternative hypothesis was to determine if the experiment would increase the ATC rate by 5% over the existing baseline of 16% for the treatment group (Refer Regression Table 3). However, we only got a 0.1% increase while an effect greater than 0.8% would have been considered to be statistically significant. (See Hypothesis section for baselines and range values.) Hence, although being assigned to ATC showed slightly larger chances of adding the product to cart, since it is not statistically significant, we failed to reject the null hypothesis that there is no difference in the ATC rate between treatment and control groups, even though we had statistical power of ~ 95% post experiment.

- **Orders ATE:**

  Regression analysis showed that orders are statistically significant at 5% level with an ATE of 0.2%. (Refer Regression Table 4) However, since order baseline rate is itself small (3-6%), to detect a 5% increase on this baseline, we needed a sample size of ~200k records to get 80% statistical power. Post treatment we only had 115k records and this gave us a statistical power of ~60% for orders outcome.

- **Orders CATE on Device**:

  Conditional on device, treatment effect CATE (Conditional Average Treatment effect) is 0.6% for PCs, 0.1% for mobile and 0.3% for tablets. Interaction terms between devices and assigned variable is not statistically significant. Hence, we fail to reject the hypothesis that the treatment produces different effects over different devices. However, for the baseline category of PCs, a 0.6% increase in order rate is a statistically significant increase over its baseline of 6%. However, eBags team had previously analyzed that order rate using PCs is almost twice as the order rate using mobile phones. Hence, this significant CATE for PCs could be because of this already existing correlation between PCs and orders, and not necessarily because of the treatment effect alone. (Refer Regression Table 4)

- **Orders CATE on New Users**:

  Conditional on new users, CATE is lesser for new users (by -0.5%) and is statistically significant. This could be because new users are negatively correlated with orders, just like PC device is positively correlated with orders. That is, new users might not right away by products and may visit the site multiple times before placing an order. Hence, it is not probably due to the treatment alone that we see a negative treatment effect for new users. (Refer Regression Table 4)

- **Orders CATE on Marketing Channels:**

  Similarly, we see a CATE of 1.1% for marketing channel baseline category of keywords. For the model with all covariates and interaction terms added, we see a CATE of 1.8% for the baseline category of recent users with PCs and marketing channel of keywords. (Refer Regression Table 4)

  Interaction between the various covariate terms and assigned variable shows further sub-group analysis and secondary experiments can be done to find causal effect conditional on these covariates.

- **Comparing Visitor vs Visit Models:**

  We found that the treatment effect coefficients were slightly more at visitor level than visitor data (ex. ATE for orders was 0.3% instead of 0.2%). This is because visitor level data provided a more wholistic view of a visitor's shopping journey at eBags and helped eliminate multiple records from analysis for each user. Even though this reduces the sample size, it helps estimate treatment effect more accurately.

- **Individual models Regression Analysis:**

  We also sub-setted data collected to analyze each of the individual models separately. Regression analysis was done with all the models data combined because it is one treatment intervention for all the 4 models and we did not want to differentiate between the various models. Individual models were showing similar results as the analysis with all the data combined (i.e. ATC not significant, orders significant). However, to get enough statistical power for these results we would need 200k records for each model (800k records in all). We currently have for each of the 4 models: Piazza ~11k, Professional Weekender ~36k, TLS ~45k and Wheeled Bag ~27k records. Definitely having more data and running the experiment for longer would help us get more statistical power for orders and for each of the individual models separately.

  All the individual models regression results can be seen in the slide presentation here:

  https://docs.google.com/presentation/d/1-40nRCCNVAx0CsvI1OdqO9fSLx9WBbqk96NkFoWlPuM/edit#slide=id.g39092c7d33_3_0

5.4 Limitations of Device Attribution:

Noncompliance:

We believe this experiment should not have any compliance issues by virtue of the fact that a user must visit one of the four selected treatment models in order to be assigned to the experiment; either in control or treatment group. Compliance with respect to treatment should automatically occur when a user in the treatment group visits the product detail page of one of the four selected models.

eBags does not specifically track if a user hovers over the product images or not in a model's PDP, i.e. no click-stream data specific to product images is available. Hence, we might not be 100% sure if the user has seen the treatment in all cases. However, it is safe to assume that there is no non-compliance because the treatment has been designed to show up in the hero (main) page of the PDP and it is generally the first thing that is noticed when a PDP loads. This has been confirmed through our initial usability pilot tests as well. Quoting from above Pilot Testing section: "Feature/benefit call-out icons needed to be on the initial main images (in addition to zoom), so that the visitor would be aware that there is more information about the product available within the images."

Attrition, and Spillover:

Attrition and the related concept of spillover is likely to occur in this experiment. Attrition would certainly occur if a subject changes devices (i.e. mobile to PC or Tablet), and/or if a subject deletes their cookies. Once attrition occurs, spillover is entirely possible as the subject will be re-randomized into either control or treatment if they visit one of the four selected models again. Hence, it is possible for a user once assigned to treatment to be assigned to control later on or vice-versa, either after deleting cookies or changing devices.

eBags team previously gathered data and found that device attrition is generally at a rate of 5-15% and users place twice as many orders from PC as mobile phones. This means we would potentially have more spillovers from mobile to PC category for orders. However, we believe this attrition is random and not differential. Everyone in general prefers to place orders using PC than mobile and there is no one specific group that might have more attrition in either control or treatment alone. Hence, we believe that since it is not differential attrition and attrition happens in both treatment and control groups at the same rate, there is no reason for us to worry about our estimates being biased because of device attribution limitations.

In the data preparation stage, we took care of removing such potential records of attrition by eliminating those users who did not view (visit) any of the models on a device but still placed orders or added-to-cart the 4 selected models on the same device. This eliminates at visitor level users who might have seen the treatment on one device but ordered or added-to-cart on another device. Co-variate balance section supports that there is minimum attrition or at least attrition that occurs proportionally within control and test.

6 Conclusion

This was a very interesting experiment and besides solidifying - through practical application - important knowledge gained in W241 Experiments and Causality, it drove meaningful business value for eBags in three areas. The three areas where eBags benefits are:

- Understanding and application of more rigorous A/B split testing protocols and practices.
- Determining a method to potentially increase orders (sales and revenue) on an average of +0.2% using the integrated product feature intervention
- Automation of feature/benefit highlighting for expansion of this initiative as well as other marketing initiatives.

In the first case, understanding and application of more rigorous A/B split testing protocols and practices, with this work eBags can re-examine its current A/B protocols and practices and evolve them to be more consistent with rigorous experimental and causality standards in business and academia. The outcome of this will undoubtedly lend confidence to learnings and decisions from A/B split testing.

In the second case, this experiment revealed that there is a potentially valuable outcome with the order ATE of +0.02% (Refer Table 4.) More power is desired for this outcome measure but the test will continue to run to attempt to reach an 80% statistical power of this conclusion. If achieved, this could improve how customers access product information and, at a minimum, could lead to other experiments that can drive this effect to a much higher and more significant treatment effect that could have a very material impact on revenue.

Finally, in developing this experiment, the MIDS W241 team developed a NLP process to extract key features and benefits from customer reviews. This could be quite valuable in automating the curation that is necessary for the treatment in this experiment (and profitability for any significant revenue effect), as well as for

other potential marketing initiatives. In short, this experiment was very valuable and successful and should definitely help eBags "bag more sales."

References

[1] Anderson, Chris. The long tail: why the future of business is selling less of more. New York: Hachette, 2014. Print.

[2] "RFC 7230 - Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing". ietf.org. Retrieved 20 August 2015.

[3] Kohavi, Ron, et. al., Ã¬Practical Guide to Controlled Experiments on the Web: Listen to Your Customers not to the HiPPO.Ã® http://ai.stanford.edu/users/ronnyk/2009controlledExperimentsOnTheWebSurvey.pdf