What works in e-commerce - a meta-analysis of 6700 online experiments

Will Browne, Mike Swarbrick Jones*

Qubit Digital Ltd

June 22, 2017

Abstract

We conduct a meta-analysis on over 6700 large e-commerce experiments, mainly from the retail and travel sectors, grouping together common treatment types performed on websites. We find that cosmetic changes have a far smaller impact on revenue than treatments grounded in behavioural psychology. This research was independently assured by PricewaterhouseCoopers UK LLP[1].

1 Introduction

Today, it is relatively simple to experiment with different versions of the same website. There are many technologies and tools that can help e-commerce businesses build and run randomised controlled trials (otherwise known as A/B tests). The amount of data available to large e-commerce sites means that businesses can measure the effect of changing design, messaging and merchandising. Over the last three years, Qubit has been helping these businesses explore which changes are associated with an increase in revenue.

In previous work [7], Qubit showed that many of the practices used in the A/B testing industry at the time were fundamentally flawed. Since its release we have seen a change in both the statistical models used in the industry, and a shift to more robust experimental procedures. In this paper, we would like to move the industry forward again, and answer the question - what kind of changes do our clients make, and how do they impact revenue?

We will present the results of a meta-analysis, conducted in 2017, on Qubit's large database of experiments. We will describe the effects of 29 treatment types and estimate the cumulative impact of these experiments on site wide revenue. The methodology used in this paper was independently assured by PricewaterhouseCoopers UK LLP (PwC)[1]. To our knowledge, this is the first published, independently assured quantitative analysis of its type. We hope it will be used to improve the quality of A/B testing, to reset expectations, and to prioritise optimisations to websites.

We have decided to separate this work into three sections to answer three slightly different questions, keeping methodologies and results together where possible. In section 2 we divide our experiments into different treatment categories, and estimate the overall impact of each of them. In section 3 we estimate the overall distribution of all experiment impacts used in this work. In section 4 we look at how A/B testing impacts overall site-wide revenue across sets of web-domains. There are a number of appendices expanding on the results of these sections.

1.1 Key findings

Due to the separated nature of this paper, and because we believe this work may be of some interest to those who are less interested in methodology, we collate some main results here.

We believe the most business-relevant metric commonly available to measure in e-commerce is *revenue per visitor*(RPV). This is the expected revenue for all visitors in an experiment

*Authors contributed equally, email: {will.browne, mike.sj}@qubit.com

[1]for full details of assured methodology and PwC assurance report please see http://www.qubit.com/sites/default/files/pdf/pwc-qubit-assurance.pdf

(including discounts, and visitors which do not purchase anything). We measure the effect a treatment has on RPV in terms of the proportional uplift e.g., increasing the mean RPV from $40 to $44 constitutes a 10% uplift. Note that this uplift only applies to those in the experiment, and does not necessarily translate to the same site-wide uplift in revenue.

We categorise roughly 2,600 experiments into a set of 29 categories, and measure a few statistics, such as the average uplift. The full list of results is in section 2.2.2. Test categories that perform best in terms of average uplift are :

- *scarcity* (stock pointers) +2.9% uplift
- *social proof* (informing users of others' behaviour) +2.3% uplift
- *urgency* (countdown timers) +1.5% uplift
- *abandonment recovery* (messaging to keep users on-site) +1.1% uplift
- *product recommendations* (suggesting other products to purchase) +0.4% uplift

Most simple UI changes to websites are ineffective. For example

- *colour* (changing the colour of elements on a website) +0.0% uplift
- *buttons* (modifying website buttons) -0.2% uplift
- *calls to action* (changing the wording on a website to be more suggestive) -0.3% uplift

We find that 90% of experiments have an effect of less than 1.2% on revenue, positive or negative (see section 3.2). However, we find that overall our clients benefit from A/B testing campaigns, some greatly (see section 4.2).

2 The effects of experiments by category

A *meta-analysis* is when one collates the data from multiple studies to identify common effects. In this section we categorise a large set of experiments into 29 treatment categories, and run a meta-analysis on each category separately. We assume that there is some simple underlying distribution for the uplift within each category, and try to estimate plausible ranges for the parameters.

2.1 Methodology

2.1.1 Qubit's A/B testing methodology

When a user enters a site running Qubit's technology, a cookie is stored in the browser which identifies the user for this and future page-views. Entry into an experiment is regulated through JavaScript that executes when targeting and segmentation criteria are met (e.g. looking at a particular page while on a mobile device).

The different treatments of the experiment are called *variants*. The visitor is randomly allocated into the control or a variant using a hash of the cookie id and the experiment id. If they are in a variant, the treatment JavaScript is injected into the page, effecting the change. Each time an experiment is served it emits events which are sent to Qubit's data processing system. The relevant events are those which indicate an experiment has been shown to a user and those which indicate a goal has been achieved. These events are aggregated then passed to a statistical model.

At Qubit, we use a Bayesian network model (the 'stats-model') to calculate beliefs about the uplifts on metrics such as conversion rate or RPV for an experiment (see figure 2.1). We will explain the structure of this model here.

Suppose we have variants labelled as $v = 0, \ldots, M$, where here we assume that 0th-index variant is the control. At Qubit we allow more than one variant as well as a control (this is sometimes called called A/B/n testing in the industry).

We divide our experiment into 'iterations' over time, labelled $i = 0, \ldots, N$, which are change points in the criteria of the experiment, for example small changes to the variant code. Keeping track of separate baseline metrics across iterations is particularly important when there is a change in the proportion of visitors allocated to each variant or the control. Baseline metrics can be different between iterations, for example conversion rate for all variants will often be higher around the time of a sale. If we change allocation during this time, some of the variants will have a higher proportion of traffic during the higher conversion rate, biasing the test. The model has been designed to account for this. This observation is of critical importance when using *multi-armed-bandit* algorithms. This is something we do

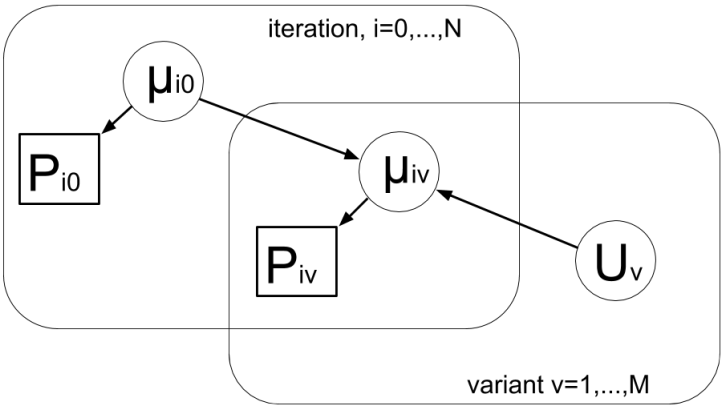

Figure 2.1: The Qubit stats model in plate notation

not believe has been well addressed in the industry, and that we would like to return to in future work.

For a variant with label $v \neq 0$, we want to measure the underlying uplift, U_v, across all iterations, which we assume is constant. Our target is the distribution $\underline{U} | \mathcal{D}$, where $\underline{U}$ is the joint distribution of the U_v's, and $\mathcal{D}$ represents all the data in the experiment. In production we impose a fairly informative prior on these uplift variables based on analysis of historical experiments, which means that the model is sceptical of large uplifts/downlifts on a test-by-test basis. Since we are aggregating experiments here, we use an uninformative prior.

We model hidden variables for the underlying mean of the metric for each iteration i and variant v, which we label as μ_{iv} (so that μ_{i0} is the mean of the metric in the control, i.e. a baseline for the metric). Let $\mathcal{D}_{iv}$ be the data for the ith iteration, vth variant. We must consider the distributions $P_{iv}(\mu_{iv}) = \mathcal{P}(\mu_{iv}|\mathcal{D}_{iv})$. If the μ_{iv}'s are modeling conversion rates, the P_{iv}'s can be modelled by binomial distributions with parameters $p = \mu_{iv}$. The situation is much more complex if the μ_{iv}'s represent means in revenue, as revenue distributions are harder to model. Some providers of e-commerce A/B testing software appear to use parametric distributions for this e.g. log-normal or exponential (e.g. [11] §10), our internal analysis has shown this can give incorrect results on real data. Revenue distributions are very discrete and multimodal, they vary wildly from business to business, and even from test to test within a single business. Instead of using a parametric model for P_{iv}, we use the Bayesian bootstrap model [9], smoothed by a kernel density estimator. Often there is a small subset of customers that spend an order of magnitude more than the average customer, in this scenario this small group can add massive uncertainty to the measurements of uplift. To counteract this, we remove the top 0.1% of customers by revenue in each test.

We have by assumption that

$$\mu_{iv} = \mu_{i0} \cdot U_v.$$

Putting all this together, we arrive at the the network in figure 2.1. We use MCMC to obtain joint samples of the U_v's.

2.1.2 Experiments included in this analysis

All data was obtained between 2014-07-09 and 2017-04-31. All experiments had revenue per visitor as a goal.

Qubit stresses to clients that they must run their experiments to a predetermined sample size. Sometimes this does not happen, for instance, if an experiment is showing a large negative expected uplift early on, the client may choose to end the experiment. To lower bias towards better performing experiments, we include experiments in our analysis that did not reach their recommended sample size, however, we do make sure that they had at least 1,000 converters in each variant and the control (the more detailed version of this rule is that some *iteration* must have at least 1,000 converters in each variant, see section 2.1.1).

All experiments were executed in a web browser. A/B testing in this domain faces a host of potential issues, for example the JavaScript that executes these experiments can fail

for a browser type in a variant. Since different browsers represent different demographics of customers, this will bias the test. An example sense-check we do at Qubit is to look at the number of visitors in the control and variant. We know the expected ratio of these two visitor groups. If the observed ratio is more than 5 standard deviations away from this, we assume that JavaScript errors have voided this experiment.

All said, these exclusion rules reduced the number of available tests to 6700.

2.1.3 Categorisation

To classify experiments, we match the recorded name, the names of variants, and the JavaScript associated with the experiments, with a set of regexes associated with each category. Each experiment may be included in multiple categories. Categories and regexes were chosen through an initial inspection of 2,000 experiment names. Categories were defined by either the visual changes made to the site, the behavioural heuristic being used, the functionality being pushed or the third party technology being tested. Combinations of categories were used to explore the effects of more complex categorisation (for example sticky navigation). For ease of reading we have put the category definitions in section 2.2.1 just before the results.

We manually checked all the title and variant names to make sure that they are correctly categorised. We took every possible measure possible to reduce bias at this stage, for example not examining the uplifts of experiments before classification. Importantly, where appropriate we make sure that the experiments were against a sensible control, i.e. not one version of a treatment vs a different version of the same treatment. This was not done for cosmetic changes, since they are nearly always just two versions of the same element on the page. For those which were not clear, we checked the past documentation written by those who conducted the test to verify, and if this was not readily available we omitted the test.

In total, roughly 3,000 tests were categorised, and we omitted about 20% through manual checking.

2.1.4 Bayesian hierarchical model for meta-analyses

Throughout this section we will assume we are only looking at one treatment, for example banners. We have a set of experimental results for this category (the joint uplift traces U_v from the stats-model MCMC process, described in section 2.1.1). We do not wish to assume that the treatment will have a uniform effect across all experiments - certainly, our clients do not assume this. Also, some experiments have more than one variant as well as the control. The uplifts of the two variants are both dependent on our measurement of the control mean, so we should not model them separately.

To combat these concerns, we use a hierarchical Bayesian network (see for example [4] §5.6). We assume our treatment has an overall true uplift for a random experiment modelled by a normal distribution with mean and variance $\mu_\mathrm{t}, \sigma_\mathrm{t}^2$. For a given experiment and variant e, v we then model our belief in the uplift U_v as

$$U_\mathrm{v} | \mu_\mathrm{t}, \sigma_\mathrm{t} \sim N(\mu_\mathrm{t}, \sigma_\mathrm{t}^2).$$

While it may seem a strong assumption for a category to be exactly normally distributed, we think it the most sensible for estimating the overall mean and variance.

If all the data in our meta-analysis is denoted $\mathcal{D}_m$, we model $U_\mathrm{v} | \mathcal{D}_m$ using the joint trace of the experiment e simulated by the stats-model in section 2.1.1, and smooth using a multi-dimensional kernel density estimator (we generally find these are well approximated by a multi-normal distribution, but we do not not make this assumption). We denote this as a potential function P_e. This completes the network shown in figure 2.2.

We give our stochastic variables μ_t, σ_t uninformative, uniform priors (for μ_t this is standard, for the variance parameters see for example [3] for justification).

We use MCMC to obtain joint samples of μ_t, σ_t. We check Gelman-Rubin statistics [6] as well as autocorrelations for convergence and mixing. We provide the mean of these parameters in the posterior in table 2.2.2, however in appendix A we also provide minimum width credible intervals containing 95% of the probability mass. For the uplift probability in table 2.2.2, we calculate the proportion of the time one would expect a sample to be positive given a particular (μ_t, σ_t), and average this across our posterior samples.

One initial concern that we had while conducting our investigation was that our clients who ran more tests would add bias to the results by running many of the same treatments on

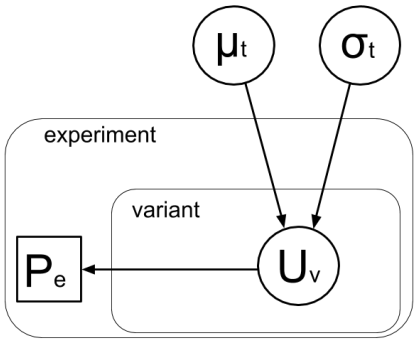

Figure 2.2: Bayesian hierarchical meta-analysis model in plate notation

their sites. If the variance between these treatments was less than the global variance, this would be problematic. To this end, we also experimented with hierarchies which modelled each client's mean for the treatment separately, as well as an 'inter-client variance'. We experimented with setting this variance to be zero, a variable shared across all clients, or variables unique to each client. We found these models did not lead to reliably better DIC scores [10], nor significantly different end results. For this reason, and to maintain model parsimony, we did not add this complexity into the model.

2.2 Results and discussion

For each category, we estimate the plausible ranges of the mean and variance of this distribution using the model described in section 2.1.4 and output the following: 1. an estimate for the mean effect (*uplift mean*) 2. an estimate for the standard deviation (*uplift s.d.*) 3. an estimate of the probability a test of this category will have a positive effect 4. an estimate of the proportion of the site's converters that a experiment of this category usually affects (*median impact*). 5. how many experiments were included in the analysis from this treatment.

Although we believe this to be the most comprehensive analysis of this type undertaken to date, there is still considerable uncertainty around the effect of some treatments. Those who are comfortable with confidence intervals/credible intervals may find the tables in appendix A more informative. We have also provided the equivalent table for the effect of these treatments on conversion rate and revenue per converter in that section.

2.2.1 Category definitions

abandonment: treatments that aim to persuade users not to leave the site after indicating abandonment behaviour.
back to top: a button used to take the user to the top of a page, usually used on mobile on longer category pages.
banner: treatments that alter or add an on-site banner.
buttons: any treatment that involves a button.
calls to action: changing the wording of copy on the page to be more suggestive, for example changing 'contact us' to 'get a quote'.
colour: any treatment that involves changing colours of elements on the page.
default setting changes: altering the default settings of site functionality. Often found on listings pages.
filters: treatments that interact with category filters such as size, colour, destination etc.
free delivery: treatments that offer or message free delivery.
image: tests involving images on the website.
landing page: treatments that are triggered only on the first page of a user's journey.
mobile navigation: altering the navigation structure for mobile.
mobile search: search box treatments, or changes to the results of a search query on mobile sites.
navigation: altering the navigation structure on a website.
nudges and pointers: tests which add additional pointers or 'tool tips' to draw a users attention to a feature.

page redesign: significant cosmetic changes that usually involve multiple elements on a page.
popup: using an image or message that pops up on screen.
product badging: adding badges to certain products to provide users with extra information (not a stock pointer).
product recommendations: treatments that recommend alternative products to users.
resizing elements: changing the dimensions of an element.
scarcity: treatments that highlight items that are low in stock, almost always by using 'stock pointers'.
search: treatments that focus on the search box, or changes to the results of a search query on site.
social proof: treatments that leverage the behaviour of other users to provide information about trending products and currently popular items.
sticky navigation: treatments which create a persistent or sticky navigation.
upsell: treatments that try to persuade the user to increase the monetary value of their basket.
urgency: treatments that use a time limit to promote urgency to complete an action before a deadline, almost always implemented using a countdown timer.
view all: treatments that default to to show all available products in a product listings page.
weather: changing content based on a the weather in the user's location.
welcome message: treatments that use a welcome message/page to introduce users to the site.

2.2.2 Results on RPV by category

treatment	uplift mean	uplift s.d.	uplift probability	median impact	number of treatments
scarcity	2.9%	2.8%	84%	38%	125
social proof	2.3%	2.5%	82%	63%	119
urgency	1.5%	2.8%	70%	36%	119
abandonment	1.1%	1.9%	71%	18%	105
product recommendations	0.4%	0.5%	76%	74%	119
welcome message	0.2%	0.6%	64%	44%	78
page redesign	0.2%	0.9%	59%	67%	83
banner	0.1%	0.3%	63%	44%	212
popup	0.0%	2.0%	50%	34%	91
colour	0.0%	0.4%	49%	81%	81
nudges and pointers	-0.0%	0.3%	48%	44%	105
resizing elements	-0.0%	1.1%	49%	85%	36
filters	-0.0%	0.9%	48%	57%	126
upsell	-0.1%	0.6%	41%	49%	99
product badging	-0.2%	0.8%	42%	64%	39
buttons	-0.2%	0.4%	33%	75%	197
image	-0.2%	0.4%	34%	40%	105
free delivery	-0.2%	1.3%	44%	50%	65
navigation	-0.2%	0.7%	35%	62%	216
search	-0.2%	0.3%	20%	60%	219
default setting changes	-0.2%	2.0%	45%	50%	58
landing page	-0.3%	0.9%	36%	39%	55
calls to action	-0.3%	0.5%	24%	71%	172
back to top	-0.4%	0.3%	12%	78%	54
view all	-0.7%	2.2%	36%	34%	30
sticky navigation	-0.7%	1.7%	32%	45%	40
mobile search	-1.0%	0.5%	5%	33%	30
weather	-1.1%	0.9%	13%	43%	27
mobile navigation	-1.7%	1.9%	17%	30%	33

2.2.3 Interesting categories

The biggest winners from our analysis all have grounding in behavioural psychology - scarcity, social proof, urgency, and to a lesser extent, abandonment recovery (see for example [2], [5]). We think that these changes alter the users' perception of the product's value. An aim of future work will be to investigate how data about users can be used to enhance the effectiveness of these approaches.

Another thing that is clear from the table is that cosmetic changes, such as changing the colour of buttons, do not constitute an effective strategy for increasing revenue. These types of changes are popular in e-commerce A/B testing, they are often easy to implement with visual editors without the need for developers. There are some high profile examples of these changes working, for example when Google ran a trial to decide which colour, out of 40 shades of blue, to use for hyperlinks on their search results page [8]. However we find the probability that these simple UI changes have meaningful impact on revenue is very low. We recommend choosing a design and sticking with it based on preference or through a qualitative process.

Personalisation is a priority for online businesses. However, scaling the effort required to implement personalisation strategies can be difficult. Within this analysis we include treatments that can be automated based on visitor context and on-site behaviour : abandonment, product recommendations, social proof, scarcity, weather and urgency. At Qubit we call these 'Programmatic experiences'. From this analysis it is clear that these are not equally effective.

It could be claimed that the negative effects for some treatments (e.g. mobile search) are due not to the treatment itself, but rather to poor implementation, for example adding flicker to a page. While we think this is a valid point, we also think that it is good to know that what kinds of changes potentially have these pitfalls, as they appear to be risky, so should be implemented more carefully, if at all.

The standard deviation of the treatments is also important to bear in mind. While some treatments, e.g. colour are neutral on average and show very little variance, others such as popups are more variant, suggesting that they can sometimes produce meaningful impact (positive and negative). In the table we have also included the *impact scores*. For each test we measured how many converters there were in each test during full days that the test was live, and divide by the total number of converters during this same time period, and give the median of this distribution. This column is primarily to remind the reader that an uplift on revenue in an experiment does not necessarily impact all users on a site, and to give a (very) rough idea of how big this discrepancy is likely to be.

2.2.4 Conversion rate vs. revenue per converter

There are two ways that website changes can influence revenue, either by changing the proportion of visitors who converted i.e. conversion rate (CR), or by changing monetary amount each converter spends; 'revenue per converter' (RPC).

One finding of this analysis is that uplifts in RPC seem to be far more neutral across the board than uplifts in conversion rate, both in terms of the size of the uplifts, as well as the variance between experiments (see appendices A.2 and A.3). There are a few exceptions - e.g. changes which are designed to upsell a customer, social proof, and product recommendations seem to be reliably positive on RPC, whereas abandonment recovery often seems to be negative (unsurprising since it is normal for abandonment messages to offer a discount as an incentive for the user to convert).

3 The distribution of all experiments

As well as measuring the impact of individual categories, we also wanted to find a distribution that describes the uplifts of all 6700 A/B tests.

3.1 Methodology

We use a kernel density estimator to approximate the distribution of all A/B tests. Measurement errors for A/B tests are equivalent to adding many convolution filters to this function

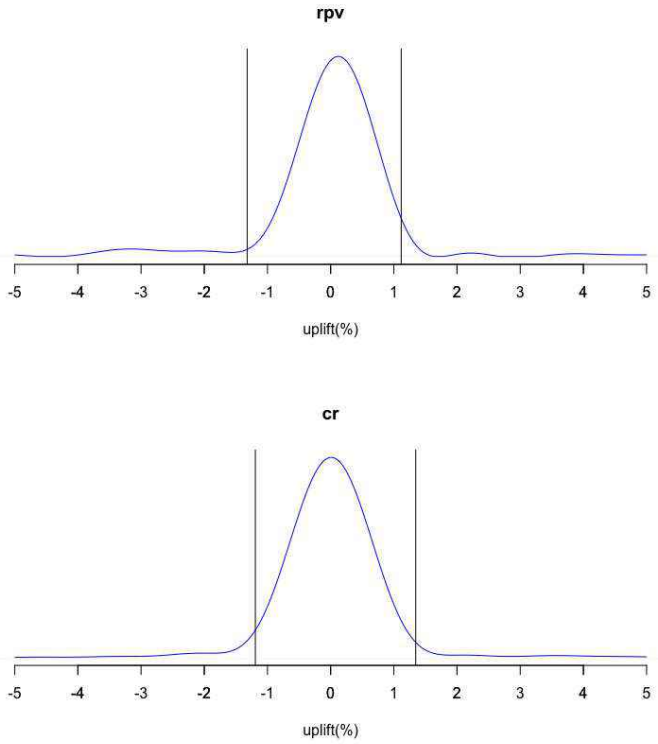

Figure 3.1: Estimated overall effects of all A/B tests

which make this problematic. This is an example of the 'deconvolution problem' (see for example [1]), which is well studied. Here we are in the regime where the measurement errors are heteroskedatic, but known. We use the R package `deconv` [12] to model this.

The deconvolution problem is notoriously hard, and here the measurement errors are large. For this reason these graphs should be taken as just a rough indication of the spread of test uplifts rather than viewed as an accurate model.

3.2 Results and discussion

We calculate graphs for the distribution of all A/B tests on revenue per visitor (RPV) and conversion rate (CR) in figure 3.1. We have marked the 5% and 95% percentiles with vertical lines. We see that the large majority of experiments have a very low effect, the distribution roughly resembles a normal distribution, but with shallow tails. This provides some rough correspondence to Sturgeon's law : '90% of everything is crud'.

Most of the treatments we measured tend to fall in the $[-1\%, 1\%]$ range for uplift. To reliably and confidently detect an uplift of 1% just on conversion rate requires about 120,000 converters (purchasing visitors) in each variant including the control. For a revenue uplift, one requires more. We will detail how one arrives at this number in appendix B. Only a small proportion of companies have enough traffic to measure uplifts of this size in a realistic time-frame.

4 Aggregation of incremental revenue over domains

Many businesses experiment primarily to learn about what kinds of changes positively impact their websites. They may increase revenue through the process of experimentation, but only if their experimentation strategy finds more positive effects than negative. We aim to measure the kind of impacts that A/B testing campaigns have on these companies revenues.

4.1 Methodology

We start with the experiments found from section 2.1.2, there were a few extra criteria used for this section.

Treatment data was obtained between 2016-10-01 and 2017-03-31 to provide a recent estimate of the current expected effect. We excluded those experiments not built by Qubit's professional services team, to limit this analysis to the team that has benefited from previous analyses of the type outlined in this paper. We only include businesses that have at least 25k converters in the 6 months under analysis.

As revealed in the results of the categorisation there are very few experiments that have a meaningful impact on RPV. As we are aggregating experiments within this analysis we are aiming to minimise the effect of aggregating lots of uncertain measurements of zero uplift. To counter this problem without biasing the data set either way, we only include experiments which have a calculated probability of uplift of above 0.9 or below 0.1.

There is a dedicated validation team at Qubit that periodically match up client's own data and the data that we record. However, in this case, we independently verify concordance of these experiments using Google Analytics data (where available) and do not use any clients for which the data from Google Analytics differs from our observation by more than 15% of the revenue observed or for where we do not have access to Google Analytics.

The results in figure 4.1 were calculated by applying the estimate of the uplift and the variance associated with this uplift to the revenue per visitor observed in the control group. This provides an incremental revenue amount per experiment.

The expected incremental revenue per property is calculated by summing the incremental revenue per experiment. The uncertainty associated with this measure is propagated by summing the variances associated with these experiments.

The estimation of the proportional impact on revenue was calculated by dividing the estimate of the total incremental revenue by the total observed revenue for that domain under the period of analysis. The uncertainty associated with this measurement was calculated by similarly scaling the standard deviations. We display the mean of the proportional uplift with a dot, as well as the 2.5th and 97.5th percentiles as the error bars.

4.2 Results and discussion

In figure 4.1 we observe businesses that do not see a significant proportional increase in site-wide revenue through experimentation. However, some businesses receive uplifts of over 5% through personalisation and experimentation strategies. The overall effect is positive.

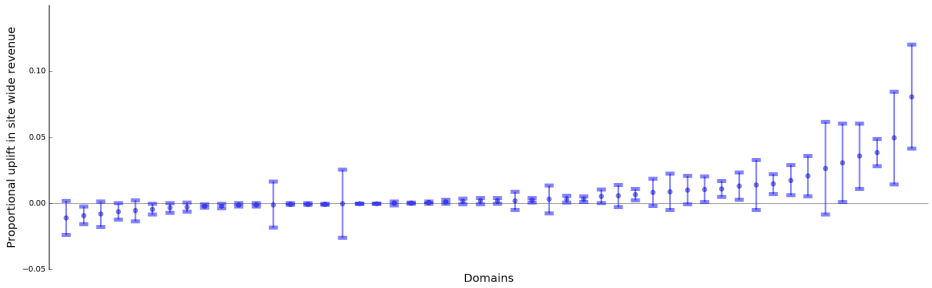

Figure 4.1: The estimated distribution of the proportional impact of an experimentation campaign on total site revenue for 50 domains over 6 months.

5 Conclusions

In some respects this analysis may make surprising reading for practitioners who try and improve revenue through online experimentation. Of the 29 common categories of treatment included in this paper only 8 have a greater than 50% probability of having a positive impact on revenue per visitor. We have found little evidence for a single treatment causing the double digit increases in revenue that we see in case studies and marketing materials. This is not to say that we see experimentation as a waste of resource, section 4.2 shows that there is significant upside available purely through experimentation. Fundamentally we believe

that learning through well designed experiments is the most powerful tool for understanding what causes measurable changes in revenue and other measures of success.

5.1 Future work

Following this analysis we would like to examine a few categories in further detail. For example - does adding a discount incentive change how well an abandonment message will perform, or does urgency cause different effects based on what the urgency is for (e.g. counting down to a sale, counting down to a delivery deadline)? Some initial analysis has shown that this is the case, but the number of tests was reduced too far for us to say anything conclusive.

Acknowledgements

This paper is the product of many years of work carried out by many people at Qubit, for which there are far too many to thank.

Prototypical versions of this analysis conducted at Qubit go back as far back as 2013, first conducted by Martin Goodson, and later by Adam Davison and the authors. The stats model was designed by Martin Goodson and Adam Davison.

This analysis would not have been possible without the tens of thousands of hours of work carried out by Qubit's professional services team and clients. We would also like to thank everyone at Qubit who was involved with this analysis in any way, particularly Jeremy Mitchell, Matthew Tamsett, Bud Goswami, Sally Zhen, Alan Clarke, Jad Sassine, Graham Cooke, Jay McCarthy, Geri Tuneva. It is also a pleasure to thank the team at PwC for their enthusiasm and diligent work in assuring our methodology.

References

[1] Delaigle, Aurore, Alexander Meister. "Density estimation with heteroscedastic error." Bernoulli (2008): 562-579. 2008

[2] Devumi., "Using Social Proof for your Digital Success", https://devumi.com/social-proof-in-digital-success/ 2015.

[3] Gelman, A., "Prior distributions for variance parameters in hierarchical models", Bayesian Anal. 1, no. 3, 515. 2006.

[4] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B,Vehtari, A., Rubin, D.B. "Bayesian Data Analysis: Third Edition", Taylor & Francis, 2013.

[5] Gupta, S. "The Psychological Effects of Perceived Scarcity on Consumers Buying Behavior.", Ph.D. Dissertation, University of Nebraska, 2013

[6] Gelman, A., Rubin, D.B. "Inference from Iterative Simulation Using Multiple Sequences", Statistical Science 7: 457511., 1992.

[7] Goodson, M. "Most Winning A/B Test Results are Illusory" http://www.qubit.com/sites/default/files/pdf/mostwinningabtestresultsareillusory_0.pdf 2013.

[8] Holson, L.M., "Putting a Bolder Face on Google", New York Times, Feb. 28, 2009.

[9] Rubin, D. B., "The Bayesian bootstrap", Annals of Statistics, 9. 130. 1981.

[10] Spiegelhalter D.J. , Best, N. G., Carlin, B. P., van der Linde, A. "Bayesian measures of model complexity and fit (with discussion)", Journal of the Royal Statistical Society, Series B. 64 (4): 583639., 2002.

[11] Stucchio, C., "Bayesian A/B Testing at VWO", https://cdn2.hubspot.net/hubfs/310840/VWO_SmartStats_technical_whitepaper.pdf 2015.

[12] Wang, X. F., Wang, B. "Deconvolution estimation in measurement error models: the R package decon." Journal of statistical software, 39(10)., 2011.

Appendices

A Detailed results tables

In this section we output detailed result tables for the mean uplifts of revenue per visitor (RPV), conversion rate (CR) and revenue per converter (RPC). This, as well as giving a point estimate for the mean and standard deviation, we also give minimum width credible intervals which cover 95% of the probability mass.

Note that while for an individual experiment, we have that RPV = CR · RPC, since these tables are based on aggregates, one should not be surprised if the results deviate from this slightly when looking at a whole category.

A.1 Revenue per visitor results

treatment	uplift mean	uplift mean CI (%)	uplift s.d.	uplift s.d. CI (%)	uplift probability
scarcity	2.9%	(2.3, 3.6)	2.8%	(2.2, 3.6)	84%
social proof	2.3%	(1.7, 2.9)	2.5%	(2.0, 3.1)	82%
urgency	1.5%	(0.7, 2.3)	2.8%	(2.0, 3.7)	70%
abandonment	1.1%	(0.4, 1.7)	1.9%	(1.3, 2.6)	71%
product recommendations	0.4%	(0.1, 0.7)	0.5%	(0.1, 1.0)	76%
welcome message	0.2%	(-0.4, 0.8)	0.6%	(0.0, 1.3)	64%
page redesign	0.2%	(-0.3, 0.7)	0.9%	(0.2, 1.7)	59%
banner	0.1%	(-0.2, 0.4)	0.3%	(0.0, 0.7)	63%
popup	0.0%	(-0.7, 0.8)	2.0%	(1.1, 2.9)	50%
colour	0.0%	(-0.5, 0.6)	0.4%	(0.0, 0.9)	49%
nudges and pointers	-0.0%	(-0.3, 0.3)	0.3%	(0.0, 0.7)	48%
resizing elements	-0.0%	(-0.8, 0.8)	1.1%	(0.0, 2.0)	49%
filters	-0.0%	(-0.4, 0.4)	0.9%	(0.5, 1.4)	48%
upsell	-0.1%	(-0.5, 0.3)	0.6%	(0.0, 1.2)	41%
product badging	-0.2%	(-1.0, 0.7)	0.8%	(0.0, 1.8)	42%
buttons	-0.2%	(-0.4, 0.1)	0.4%	(0.0, 0.8)	33%
image	-0.2%	(-0.6, 0.2)	0.4%	(0.0, 1.0)	34%
free delivery	-0.2%	(-0.8, 0.4)	1.3%	(0.6, 2.0)	44%
navigation	-0.2%	(-0.5, 0.1)	0.7%	(0.0, 1.2)	35%
search	-0.2%	(-0.5, 0.1)	0.3%	(0.0, 0.7)	20%
default setting changes	-0.2%	(-1.1, 0.6)	2.0%	(1.3, 2.8)	45%
landing page	-0.3%	(-0.9, 0.3)	0.9%	(0.1, 1.6)	36%
calls to action	-0.3%	(-0.6, 0.0)	0.5%	(0.0, 0.9)	24%
back to top	-0.4%	(-0.8,-0.0)	0.3%	(0.0, 0.7)	12%
view all	-0.7%	(-2.0, 0.5)	2.2%	(0.7, 3.6)	36%
sticky navigation	-0.7%	(-1.7, 0.2)	1.7%	(0.1, 3.0)	32%
mobile search	-1.0%	(-1.7,-0.3)	0.5%	(0.0, 1.1)	5%
weather	-1.1%	(-2.1,-0.0)	0.9%	(0.0, 2.1)	13%
mobile navigation	-1.7%	(-2.9,-0.5)	1.9%	(0.1, 3.3)	17%

A.2 Conversion rate results

treatment	uplift mean	uplift mean CI (%)	uplift s.d.	uplift s.d. CI (%)	uplift probability
scarcity	2.9%	(2.3, 3.5)	2.9%	(2.4, 3.5)	83%
social proof	1.9%	(1.4, 2.4)	2.2%	(1.8, 2.6)	79%
abandonment	1.6%	(1.1, 2.0)	1.8%	(1.4, 2.3)	80%
urgency	1.5%	(0.9, 2.1)	2.3%	(1.8, 2.9)	74%
welcome message	0.5%	(-0.2, 1.3)	2.7%	(2.0, 3.6)	57%
resizing elements	0.2%	(-0.3, 0.8)	1.1%	(0.5, 1.7)	59%
product badging	0.2%	(-0.3, 0.7)	0.4%	(0.0, 1.0)	69%
page redesign	0.1%	(-0.2, 0.4)	0.6%	(0.2, 1.0)	59%
product recommendations	0.0%	(-0.2, 0.2)	0.4%	(0.2, 0.7)	52%
free delivery	0.0%	(-0.5, 0.5)	1.4%	(1.0, 1.9)	50%
buttons	-0.0%	(-0.2, 0.1)	0.3%	(0.0, 0.6)	46%
banner	-0.1%	(-0.2, 0.0)	0.1%	(0.0, 0.2)	24%
default setting changes	-0.1%	(-0.7, 0.5)	1.5%	(1.1, 2.0)	47%
filters	-0.1%	(-0.3, 0.1)	0.6%	(0.3, 0.8)	42%
navigation	-0.1%	(-0.3, 0.1)	0.2%	(0.0, 0.5)	29%
image	-0.2%	(-0.5, 0.1)	0.7%	(0.3, 1.1)	40%
popup	-0.2%	(-0.6, 0.3)	1.6%	(1.0, 2.2)	45%
back to top	-0.2%	(-0.4, 0.1)	0.3%	(0.0, 0.6)	24%
colour	-0.2%	(-0.5, 0.1)	0.3%	(0.0, 0.7)	28%
calls to action	-0.2%	(-0.4, 0.0)	0.5%	(0.3, 0.7)	35%
upsell	-0.3%	(-0.5,-0.0)	0.3%	(0.0, 0.6)	19%
search	-0.3%	(-0.5,-0.1)	0.2%	(0.0, 0.5)	13%
landing page	-0.3%	(-0.8, 0.1)	0.9%	(0.3, 1.7)	36%
nudges and pointers	-0.3%	(-0.8, 0.1)	1.7%	(1.4, 2.0)	42%
view all	-0.4%	(-1.3, 0.5)	1.6%	(0.6, 2.7)	40%
sticky navigation	-0.5%	(-1.1, 0.0)	0.8%	(0.0, 1.6)	23%
mobile search	-0.7%	(-1.2,-0.2)	0.5%	(0.0, 1.1)	10%
weather	-1.0%	(-1.8,-0.2)	1.0%	(0.0, 2.0)	16%
mobile navigation	-1.1%	(-1.9,-0.3)	1.4%	(0.6, 2.3)	21%

A.3 Revenue per converter results

treatment	uplift mean	uplift mean CI (%)	uplift s.d.	uplift s.d. CI (%)	uplift probability
upsell	0.5%	(0.1, 0.9)	0.8%	(0.2, 1.3)	74%
social proof	0.3%	(0.1, 0.6)	0.4%	(0.0, 0.8)	81%
product recommendations	0.3%	(0.1, 0.5)	0.2%	(0.0, 0.4)	91%
colour	0.2%	(-0.2, 0.6)	0.3%	(0.0, 0.7)	76%
nudges and pointers	0.2%	(-0.1, 0.4)	0.2%	(0.0, 0.4)	81%
scarcity	0.2%	(-0.1, 0.5)	0.2%	(0.0, 0.5)	74%
page redesign	0.1%	(-0.2, 0.5)	0.3%	(0.0, 0.6)	68%
search	0.1%	(-0.1, 0.3)	0.2%	(0.0, 0.5)	69%
sticky navigation	0.1%	(-0.5, 0.6)	0.6%	(0.0, 1.3)	57%
landing page	0.1%	(-0.3, 0.4)	0.2%	(0.0, 0.6)	61%
banner	0.1%	(-0.2, 0.3)	0.2%	(0.0, 0.5)	59%
filters	0.0%	(-0.3, 0.3)	0.8%	(0.6, 1.1)	50%
popup	0.0%	(-0.4, 0.4)	0.3%	(0.0, 0.8)	50%
navigation	-0.0%	(-0.2, 0.2)	0.3%	(0.0, 0.7)	48%
image	-0.0%	(-0.3, 0.3)	0.2%	(0.0, 0.5)	46%
product badging	-0.0%	(-0.7, 0.6)	0.5%	(0.0, 1.1)	47%
urgency	-0.1%	(-0.4, 0.3)	0.4%	(0.0, 0.9)	43%
calls to action	-0.1%	(-0.3, 0.2)	0.2%	(0.0, 0.5)	36%
weather	-0.1%	(-0.8, 0.7)	0.4%	(0.0, 1.0)	44%
view all	-0.1%	(-0.8, 0.6)	0.8%	(0.1, 1.5)	44%
default setting changes	-0.1%	(-0.5, 0.3)	0.5%	(0.0, 1.1)	38%
free delivery	-0.2%	(-0.7, 0.4)	1.4%	(0.9, 2.0)	45%
buttons	-0.2%	(-0.4, 0.0)	0.2%	(0.0, 0.5)	20%
back to top	-0.3%	(-0.6, 0.0)	0.2%	(0.0, 0.5)	15%
mobile search	-0.3%	(-0.8, 0.2)	0.4%	(0.0, 1.0)	21%
mobile navigation	-0.4%	(-0.9, 0.2)	0.5%	(0.0, 1.2)	24%
welcome message	-0.4%	(-1.1, 0.3)	2.1%	(1.5, 2.8)	42%
resizing elements	-0.4%	(-1.0, 0.1)	0.4%	(0.0, 0.9)	18%
abandonment	-0.6%	(-1.1,-0.2)	1.1%	(0.6, 1.5)	27%

B How much data A/B tests need

At Qubit, we recommend clients run experiments with a one-tailed false positive rate of 5%. Said another way, we accept a test as a 'winner' if we believe the probability of uplift is greater than 95%. The sample size required is such that it that allows one to detect a 5% uplift with 80% probability (that is, if the genuine underlying uplift is 5%, the test has an 80% chance of winning). We say that 5% is the target uplift, and 80% is the power.

These numbers are industry standard. If only looking for an uplift in conversion rate, this typically requires a sample about 5,700 *converters* in both the variant and the control (for revenue experiments, there is no simple number for this, but it is generally at least twice as large). Keeping the significance and power the same, but varying the target uplift away from 5%, the sample size follows roughly an inverse square law :

$$\text{sample size} \propto (\text{target uplift})^c,$$

where $c \approx -1.9$. So to reliably detect an uplift of size 1%, one would need more than 20 times the amount of data as with the 5% target uplift : roughly 120,000 converters in the control and each variant.

C In depth statistics for treatments

In this section we will show detailed statistics for the treatment categories in the meta-analysis. For brevity we have only included the plots mentioned by name in section 2.2. For the full list, and higher resolution images see `https://github.com/mikesjqubit/qubit-meta-analysis-results`. For each treatment we output four graphs. The first two are summary statistics for every experiment used for revenue per converter (RPV) and conversion rate (CR) - we display the mean of the experiment with a dot, as well as the 5th and 95th percentiles as the error bars.

The second pair of graphs is a detailed summary of the posterior distributions output from the meta-analysis. We provide 2-d scatter plots for RPV and CR, as well as contours marking the 90th and 99th percentiles. These were generated by running a grid search over the plausible range of the variables and using the potential functions from section 2.1.4.

C.1 abandonment

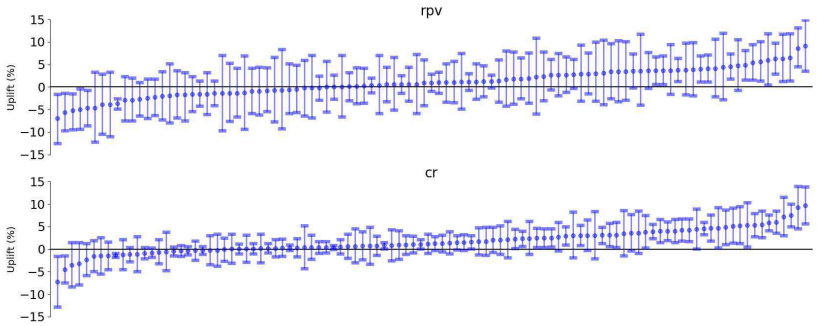

Figure C.1: 'abandonment' test summaries

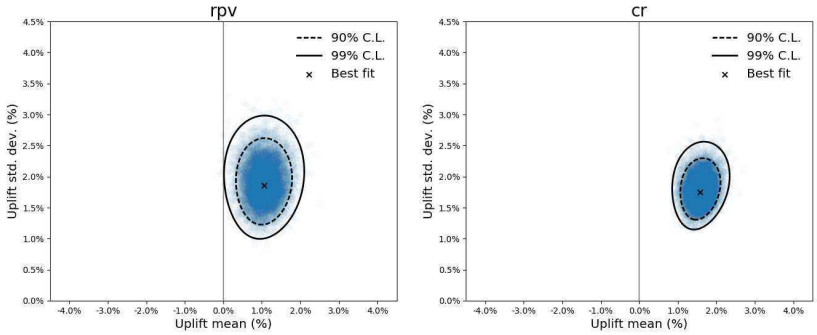

Figure C.2: 'abandonment' posterior density plots

C.2 buttons

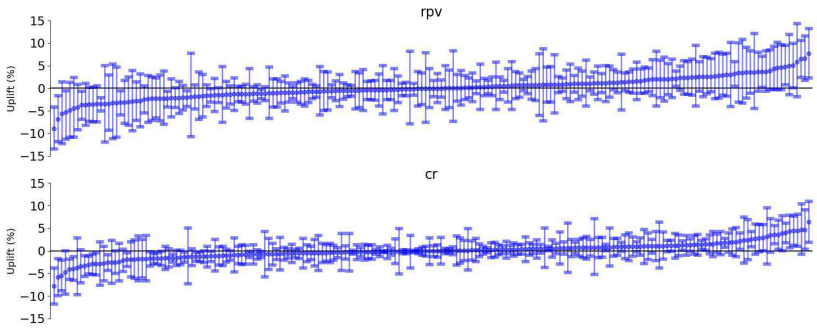

Figure C.3: 'buttons' test summaries

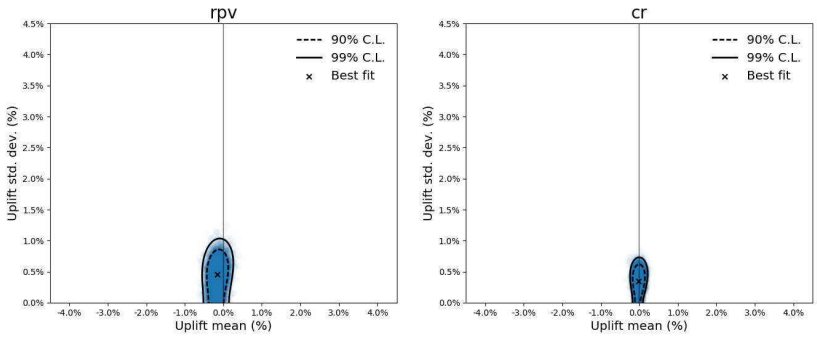

Figure C.4: 'buttons' posterior density plots

C.3 colour

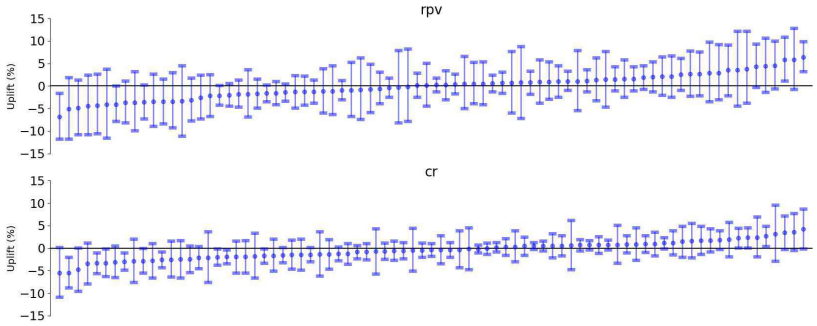

Figure C.5: 'colour' test summaries

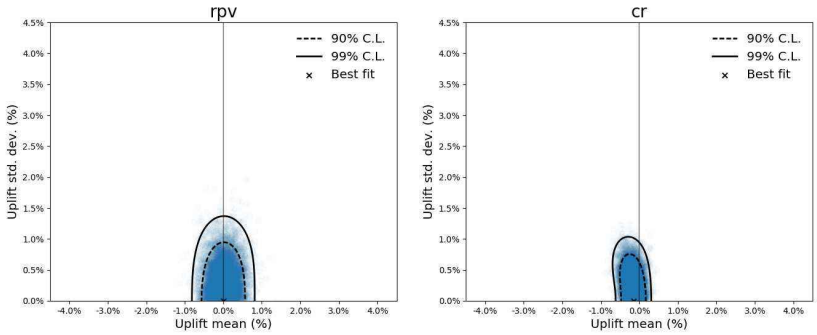

Figure C.6: 'colour' posterior density plots

C.4 calls to action

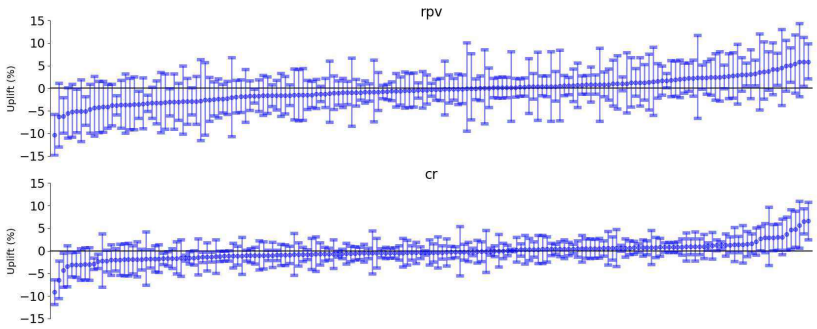

Figure C.7: 'calls to action' test summaries

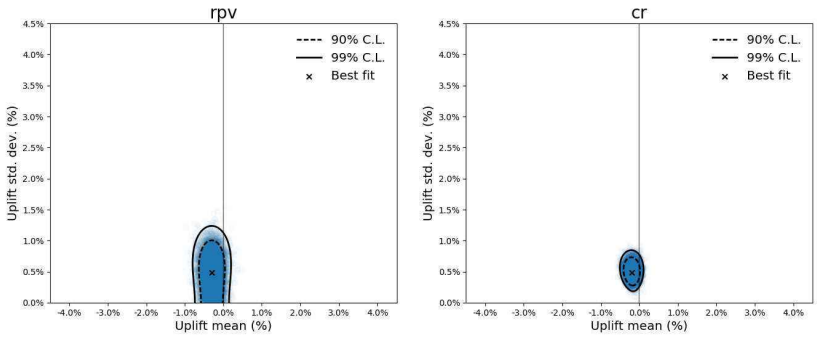

Figure C.8: 'calls to action' posterior density plots

C.5 mobile navigation

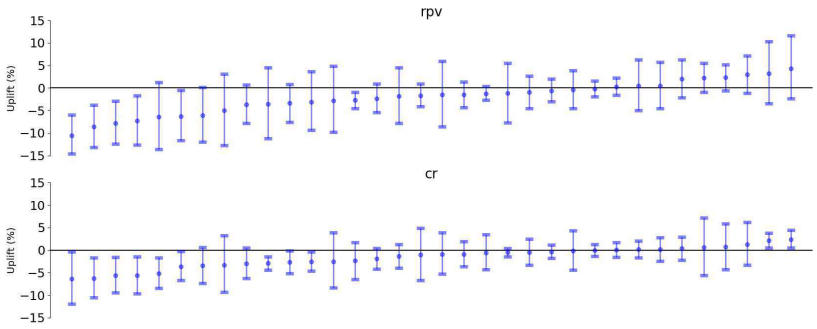

Figure C.9: 'mobile navigation' test summaries

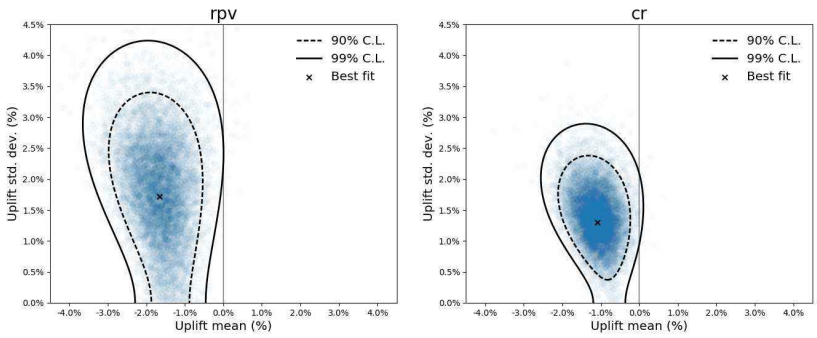

Figure C.10: 'mobile navigation' posterior density plots

C.6 product recommendations

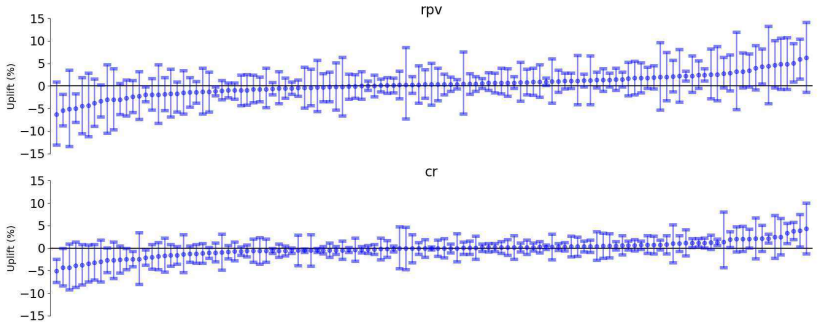

Figure C.11: 'product recommendations' test summaries

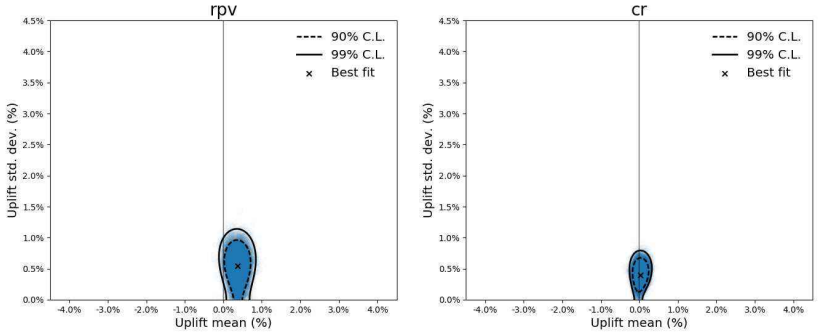

Figure C.12: 'product recommendations' posterior density plots

C.7 scarcity

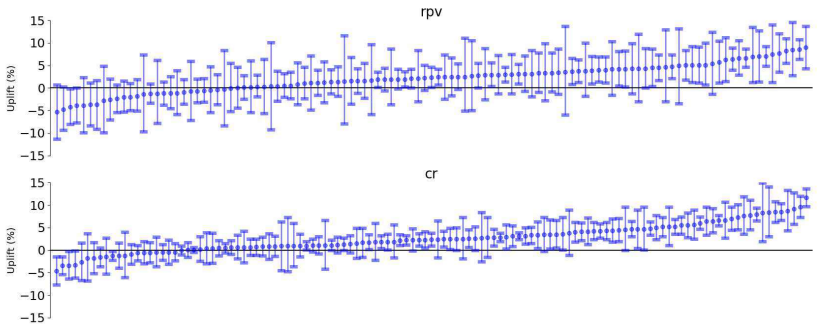

Figure C.13: 'scarcity' test summaries

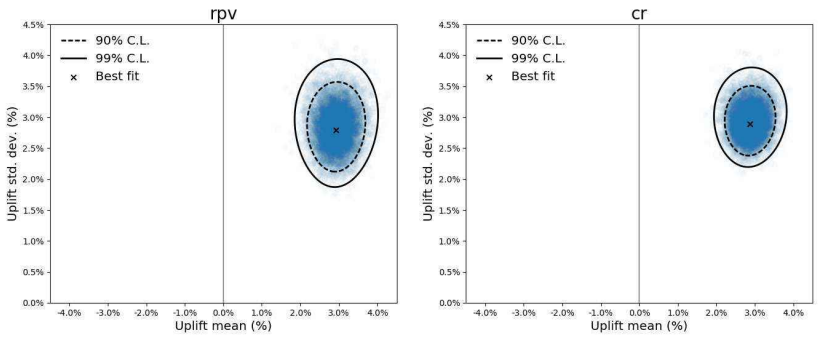

Figure C.14: 'scarcity' posterior density plots

C.8 social proof

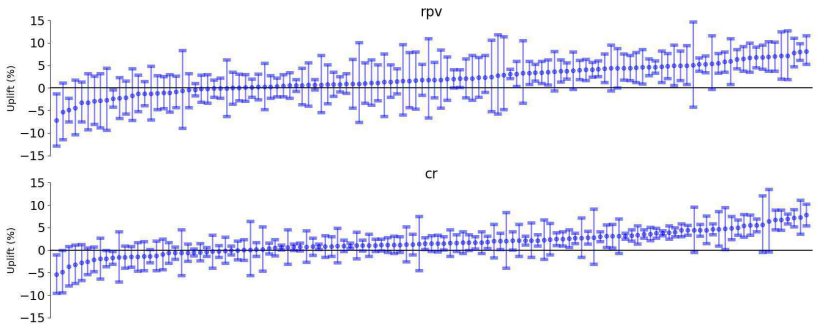

Figure C.15: 'social proof' test summaries

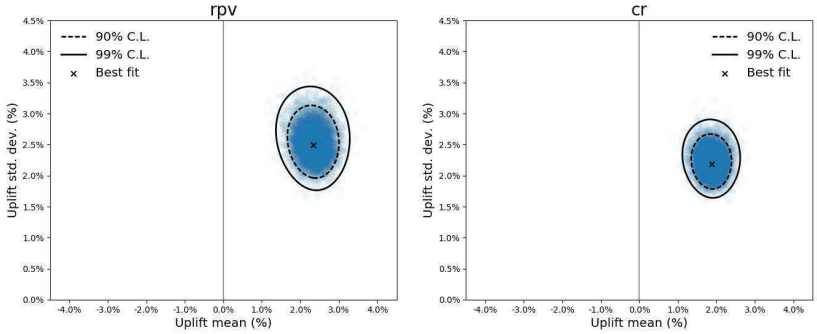

Figure C.16: 'social proof' posterior density plots

C.9 upsell

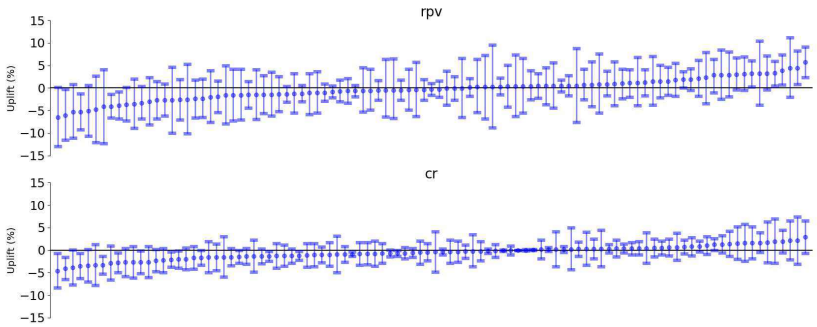

Figure C.17: 'upsell' test summaries

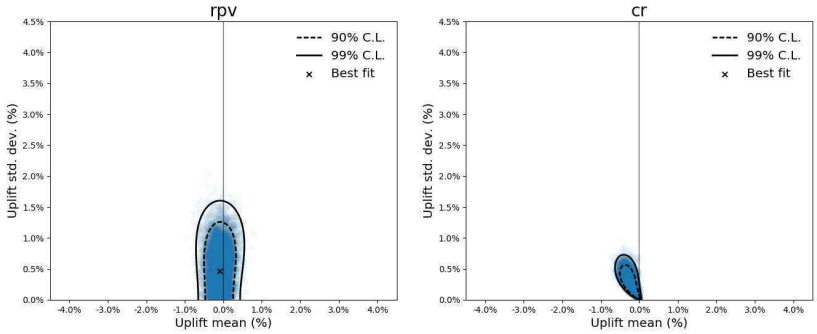

Figure C.18: 'upsell' posterior density plots

C.10 urgency

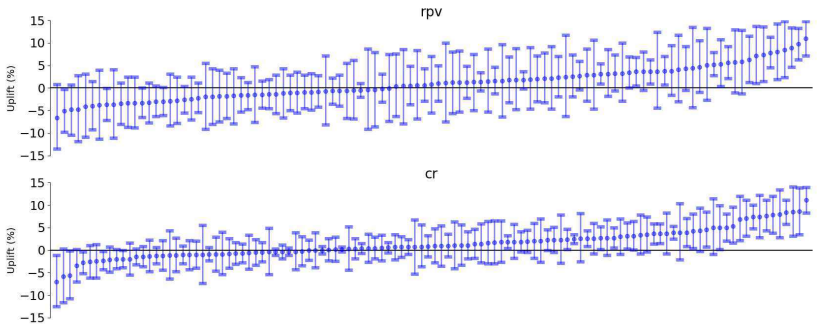

Figure C.19: ‘urgency’ test summaries

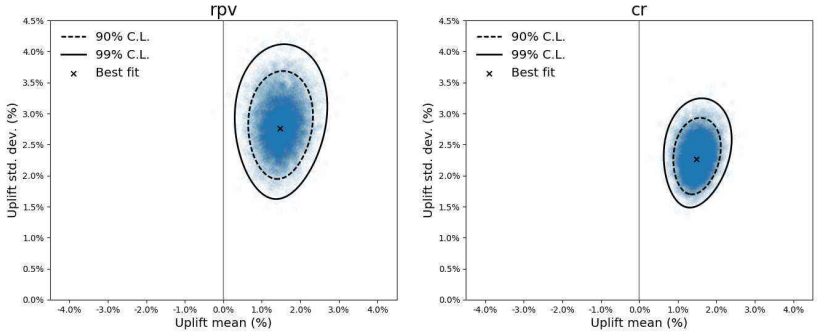

Figure C.20: ’urgency’ posterior density plots

C.11 weather

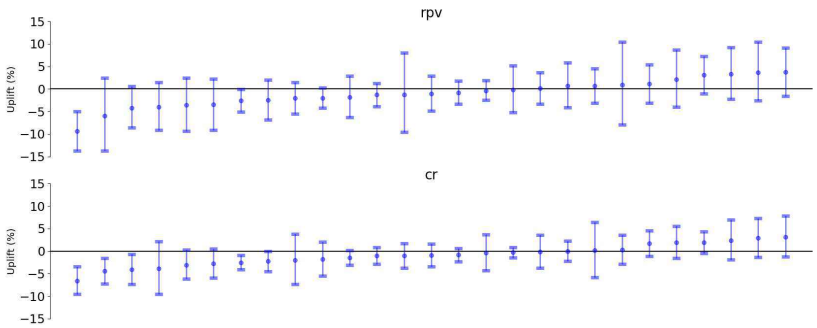

Figure C.21: ‘weather’ test summaries

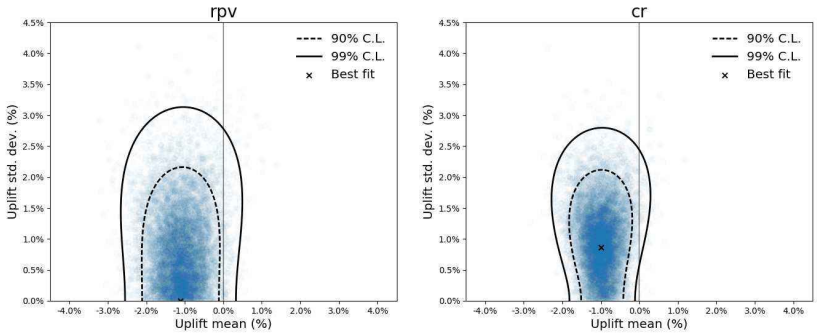

Figure C.22: ’weather’ posterior density plots

D Segmented vs Non-Segmented treatments

Treatments can be exposed to all users or they can be targeted at specific sub-groups. One assumption implicit in the rise of personalisation technology is that the more targeted a treatment is to specific group, the more effective it will be. We define a segmented experiment as any experiment that uses the Qubit segmentation technology as a criteria for the user to be included in the experiment. By analysing the expected effect of segmented treatments vs non-segmented treatments we observe a nearly 3 fold difference in expected uplift in revenue. Interestingly, we also observe a much larger standard deviation for the effect of segmented experiments (see figures D.1 and D.2). As personalisation increases in importance segmentation will be a crucial tool. A challenge for businesses attempting to value the impact of these more targeted approaches is that the smaller the group of users targeted the harder it will be to measure the impact of the changes made. Future analysis of this type may require different approaches to prove the revenue impact of highly personalised treatments.

D.1 segmented

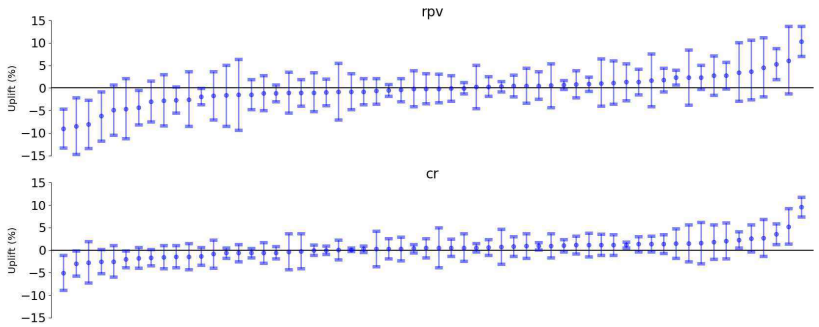

Figure D.1: 'segmented' test summaries

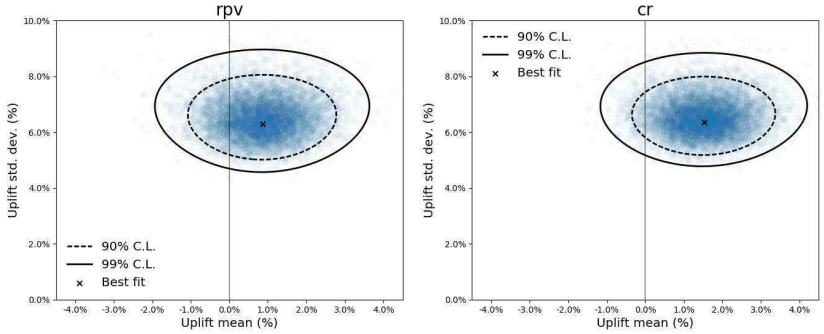

Figure D.2: 'segmented' posterior density plots

D.2 unsegmented

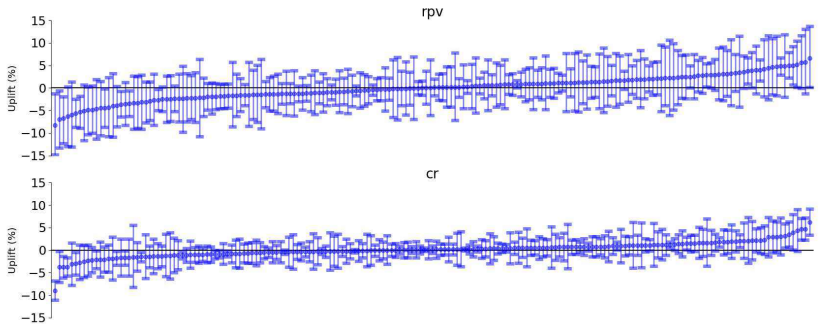

Figure D.3: 'unsegmented' test summaries

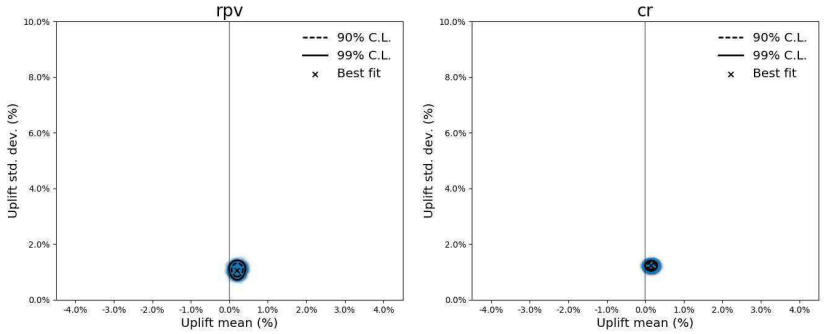

Figure D.4: 'unsegmented' posterior density plots