Why Marketplace Experimentation Is Harder than it Seems: The Role of Test-Control Interference

THOMAS BLAKE, eBay Research Labs
DOMINIC COEY, eBay Research Labs

Classical statistical inference of experimental data assumes that the treatment affects the test group but not the control group. This assumption will typically be violated when experimenting in marketplaces because of general equilibrium effects: changing test demand affects the supply available to the control group. We illustrate this with an email marketing campaign performed by eBay. Ignoring test-control interference leads to estimates of the campaign's effectiveness which are too large by a factor of around two. We present the simple economics of this bias in a supply and demand framework, showing that the bias is larger in magnitude where there is more inelastic supply, and is positive if demand is elastic.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics; G.3 [**Probability and Statistics**]: Experimental Design; K.4.4 [**Computers and Society**]: Electronic Commerce

General Terms: Experimentation, Marketplace

Additional Key Words and Phrases: Online marketplace, auctions, AB testing, marketing email tests

1. INTRODUCTION

Experimentation, or A/B testing, is an increasingly popular tool for guiding product decisions online [Crook et al. 2009; Kohavi et al. 2009]. In typical practice, some users are assigned to a product experience different from the default. The difference may be in design, content, or some other dimension of user experience. Many of the issues involved in analyzing data generated by such experiments are well understood.[1] We focus on a concern which has received less attention in the literature: the control group may be indirectly affected by the treatment because they interact in a marketplace.

Online marketplaces are by their very nature interconnected. A new marketing campaign, product design, or search algorithm available to some users may change their demand, and consequently change the supply available to other users. The treatment affects not only the targeted users, but also others in the market. Classical statistical inference from experimental data assumes that the treatment affects the test group but not the control group. The *stable unit treatment value* assumption is violated in the case of such test-control "interference" [e.g. Cox 1958; Rubin 1974, 1986].[2] If outcomes in both groups change as

Author's addresses: T. Blake, thblake@ebay.com, and D. Coey, dcoey@ebay.com. Both at eBay Research Labs, 2065 Hamilton Ave San Jose, CA 95125.

EC'14, June 8–12, 2014, Stanford, CA, USA.

http://dx.doi.org/10.1145/2600057.2602837

[1] For example, the signal-to-noise ratio may be very small, especially in online advertising [Lewis and Rao 2013]; results may be incorrectly judged to be significant with multiple comparisons [Dunn 1961; Benjamini and Hochberg 1995], or if the experiment is stopped on the basis of the data generated so far [Bassler et al. 2008; Simmons et al. 2011]; and with multiple treatments, treatment effects may interact [Fisher 1935; Montgomery 1984].

[2] Kershner and Federer [1981] and David and Kempton [1996] relax this assumption in various non-market settings, including interference within-person and across time, and interference across plots in agricultural ex-

a result of the experiment, a simple test-control comparison will in general be a biased estimate of the true effect on market outcomes.

Consider the example of testing a new search engine ranking algorithm which steers test buyers towards a particular class of items for sale. If test users buy up those items, the supply available to the control users declines. Thus sales to the control group change too. This causes the test-control comparison to be biased in the following sense: the difference between test and control users' sales when the test receives the treatment is not the same as the difference between *all* users' sales with and without the test receiving the treatment.

Some of the experimental literature in online marketplaces has acknowledged the potential for this kind of bias. Lucking-Reiley [1999], Reiley [2006], and Einav et al. [2011] describe how test-control interference could potentially arise in the context of seller experiments. Blake et al. [2013] restrict their analysis of a marketing experiment to non-auction sales on eBay to limit the effect of supply constraints that would cause test-control interference.[3] In a simulation of an online market for short-term lodging, Fradkin [2013] finds that user level search experiments can overstate market-wide outcomes by 90%. Kohavi et al. [2009] notes that auction level experiments might be preferable to user randomization for marketplaces like eBay, an approach we explore empirically.

We make two contributions to the existing literature. First, we present what is to our knowledge the first empirical evidence of an experiment in which there was significant test-control interference generated by users interacting in a marketplace.[4] Second, we analyze the simple economics of market experiments in a supply and demand framework, show why this interference is likely to arise quite generally, and show how the bias depends on supply and demand elasticities.

The empirical example we study is an email marketing campaign experiment run by eBay. In this experiment, users were randomized into receiving marketing emails or not. Test and control users interact within auctions, however, because if the email induces a test user to win an auction, it may also induce a control user to lose. To avoid the bias generated by these within-auction user interactions, we compare auctions with many test users and few control users to those auctions with few test and many control users. Our empirical strategy finds that the effect of the campaign is not significantly different from zero, whereas a naive user comparison would suggest that the campaign was very successful. We also find that control users' revenue is lower when they compete against test users than control users, as would be expected with test-control interference. Accordingly, we offer this experiment as a cautionary tale in the evaluation of online marketplace experiments.

We then present the simple economics of marketplace experiments graphically in a supply and demand framework. We derive formally how the bias from a test-control comparison relates to supply and demand elasticities. The magnitude of the bias increases as supply becomes more inelastic, and will be non-zero unless supply is perfectly elastic or demand is unit elastic. The direction of the bias depends on the elasticity of demand: it will be positive when demand is elastic, and negative otherwise.

---

periments. Rosenbaum [2007] observes that Fisher's randomization test of the null of no effect has the correct level even with interference. He shows how to construct confidence intervals on the effect of the treatment on the test *relative to the control* with interference, but notes that this will in general be different from the overall effect of the treatment on test and control. Sinclair et al. [2012] assess the stable unit treatment value assumption in multi-level experiments.

[3] Although Brown et al. [2010] do not explicitly mention test-control interference, they implicitly address it by supplementing findings from a randomized trial with a market-wide natural experiment.

[4] Fradkin [2013], by contrast, simulates a hypothetical experiment with a calibrated matching model.

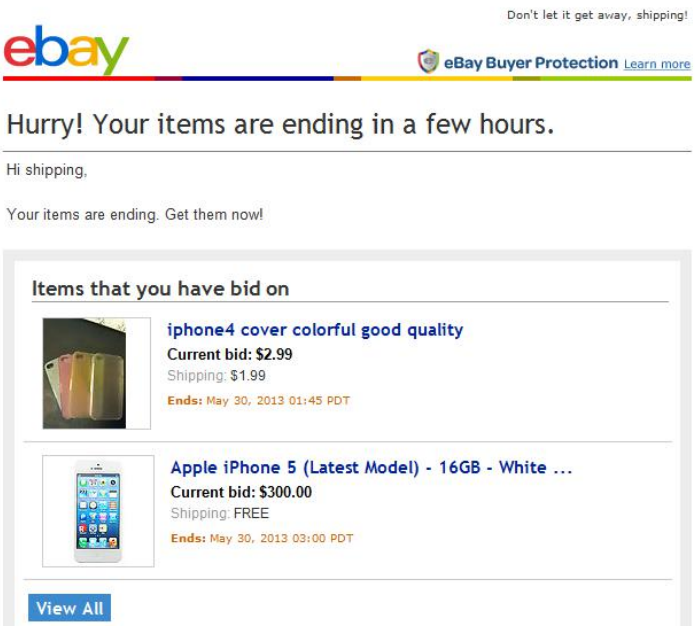

Fig. 1. Sample Email Received by Test Group

2. A CAUTIONARY TALE: TEST-CONTROL INTERFERENCE IN ONLINE AUCTIONS

eBay's "Bid Item Ending Soon" experiment ran for four weeks in July and August 2013, targeting users who placed bids in auctions during this period. 4.9 million U.S. eBay users qualified for the experiment. Users were randomized with equal probabilities to test and control groups. Approximately six hours before the end of any auction, test users were sent an email if they had previously bid in that auction. The email served to remind bidders that their auctions were "ending soon", and encouraged them to try to win. Figure 1 displays a sample email. Control users did not receive this email. This test is typical of online experiments where a subset of users receives a treatment and is compared to the remaining users. Auction revenue is the outcome of interest.[5]

A simple example clarifies why comparing test and control users' revenue can give an inaccurate impression of the effect of the treatment. Table I shows bids in a second-price auction by two bidders, A and B, in two scenarios: i) no one receives an email, and ii) only user A does, in which case user A is the test user and user B is the control user. In the first case user B bids 100, user A bids 99, and B wins at a price of 99. In the second case we assume that the email persuades user A to increase his bid to 101 and does not affect user B's bid, so that A wins at a price of 100.

When the email is sent out, the difference between test and control revenue (i.e. between user A's revenue and user B's revenue) is 100. But the true effect of the email on revenue is the difference in total revenue in the with- and without-email cases–a much smaller difference of just 100 - 99 = 1. The test-control interference here is being generated by the fact that if user A wins the auction, user B must lose. Comparing test and control sales captures the incremental revenue generating effect of the email. But it also captures the revenue reassignment effect: revenue which would have been generated anyway, by user

[5]We follow the convention in the auctions literature and use "revenue" to refer to the selling price of the good, rather than the revenue earned from the sale by eBay.

Table I. How Test-Control Interference Arises

	No Email Sent	Only User A Receives an Email
User A's Bid	99	101
User B's Bid	100	100
Revenue from User A	0	100
Revenue from User B	99	0
Total Revenue	99	100

B buying the good, is being erroneously attributed to the email, because now user A is buying the good.

With fixed supply, a treatment which makes the test user more likely to win must therefore make the control user more likely to lose, and the usual methodology of comparing test and control individuals' revenue gives a biased estimate of the true treatment effect. In this unit supply and unit demand case example, and in our empirical analysis of the email experiment, the naive test-control comparison overstates the true effect of the experiment. This is not always true however, and Section 4 shows how the magnitude and sign of the bias depends on supply and demand elasticities.

If the experiment were conducted at the auction level, with the email being sent to all participants in a randomly selected group of test auctions, this within-auction revenue reassignment bias could not arise. Our empirical strategy seeks to imitate this more sound experimental design with the data from the original, flawed experiment. The basic intuition is simple: we can compare auctions that, by chance, had all test bidders to those auctions that, by chance, had all control bidders. That is, we can compute both the naive user level estimates, and the less biased auction level estimates, from the original user level experiment.

3. EMPIRICAL ANALYSIS OF EMAIL EXPERIMENT

3.1. Summary Statistics and Regression Results

Table II shows summary statistics on the 10.4 million auctions in which a test or control bidder entered during the experiment window. Some users are in neither the test nor the control group because they have opted out of receiving emails. Auction revenue is winsorised at the 99.9th percentile to reduce the influence of outliers.

Table II. Auction Participation by Bidder Group, Number of Bids, and Revenue

N	Total Bidders	Control Bidders	Test Bidders	Other Bidders	Bids	Revenue ($)
10,425,390	3.75	1.39	1.40	0.96	6.99	33.87
	(2.83)	(1.36)	(1.36)	(1.30)	(7.26)	(93.41)

Table shows means and standard deviations of the number of auction participants by various bidder groups, of the total number of auction bids, and of auction revenue, for the auctions during the experiment period with some test or control entrants. Some users opt out of emails, hence the "Other Bidders" group. Auction revenue is winsorised at the 99.9th percentile.

Table III shows the mean normalized revenue from the test and control groups over the experiment period, along with standard errors.[6] The test group buys significantly more than the control group. It appears as if sending the email has generated roughly 0.74% in extra revenue. This is a quite substantial difference: for any online retailer several such campaigns would make for a banner year. Given that the marginal cost of sending emails

[6]We calculate standard errors assuming revenue and total bids are independent across users, ignoring dependence induced by users participating in the same auctions. This gives a realistic impression of the standard industry "naive" user level analysis.

is close to zero, and ignoring the bias caused by test-control interference, this would imply a large, positive return on investment.[7]

Table III. Normalized Revenue, Test and Control Bidders

	Control	Test	Difference
N	2,427,629	2,428,500	
Revenue	100.00	100.74	0.74
	(0.20)	(0.21)	(0.29)
Total Bids	100.00	100.76	0.76
	(0.24)	(0.25)	(0.34)

Table shows mean revenue per user and mean number of total bids per user in the control and test groups as well as their differences, with standard errors in parentheses. Revenue per user and total bids per user are winsorised at the 99.9th percentile. Mean revenue per user and mean number of total bids per user are normalized so that the means in the control group are 100.

We now investigate whether these test-control differences are mostly because of truly incremental revenue, or rather reflect revenue being "reassigned" from control to test groups. A better experimental design would randomize auctions, rather than individuals, into test and control, with all participants in test auctions and no participants in control auctions receiving the email. This would not be subject to the revenue reassignment bias described above, as test and control users would never compete in the same auctions. Our empirical strategy mimics this design, exploiting exogenous variation in email intensity across auctions, rather than across individuals. Bidders are randomly assigned to test or control, and some auctions end up with more test bidders than others. We measure the effect of the email by comparing outcomes in auctions with few test entrants to those with many test entrants, holding fixed the total number of test and control entrants. This exploits the fact that while the number of entrants is endogenous, the proportion that is test or control is exogenous.

Figure 2 shows, for auctions with between one and five test and control entrants, how revenue changes with the number of test bidders.[8] The various subplots group auctions by the number of test or control bidders combined that were eligible for the treatment. For instance, the first plot represents all auctions with only one test or control bidder combined. For each set of auctions, the vertical axis plots the revenue per auction and the horizontal axis plots the number of users that are in the test group. Blue dots represent mean revenue, and red dots are 95% confidence intervals for mean revenue. If the email campaign generated detectable incremental revenue, revenue should tend to increase with the number of test bidders in each graph.[9] Instead, there does not appear to be any trend in revenue. In particular, there is no significant difference in revenue between the cases when all test and control users are test, and when all test and control users are control (Online Appendix Figure 7 plots these differences and their standard errors).

Auction level regressions also support this conclusion. We regress auction revenue and the total number of bids in the auction on the number of test bidders in that auction, with fixed effects for the combined number of test and control bidders, and for the number of other bidders. These regressions serve both to pool effects across auctions with different

[7] An extra email campaign might decrease customers' response to existing email campaigns. This effect would be reflected in the revenue figures in Table III, as those numbers include all purchases made by the test and control groups.

[8] In the Online Appendix, Figure 5 shows the analogous plot for bids, and Figure 6 shows the distribution of the number of test bidders per auction for auctions with between one and five test and control entrants.

[9] This is true even if there is only one test or control bidder. Users who opt out of emails are not included in either the test or control groups. There may thus be multiple bidders in auctions with one test or control bidder.

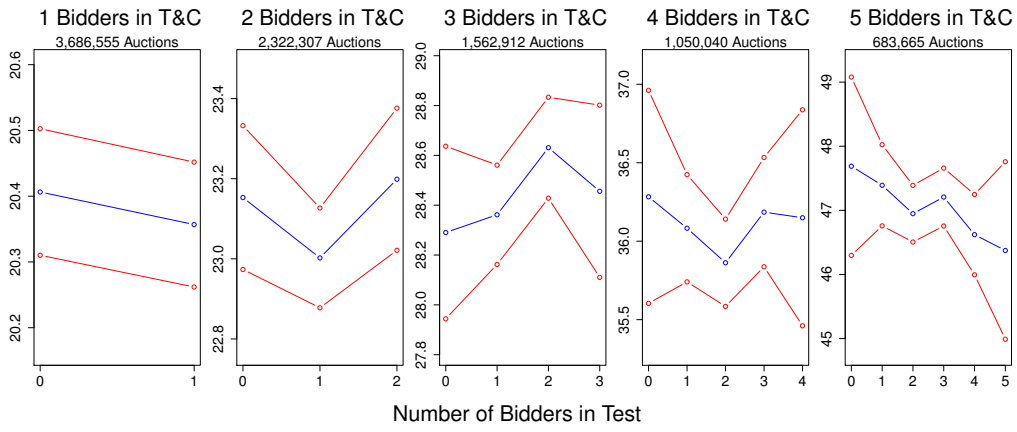

Fig. 2. The Effect of Emails on Revenue ($)

Plots show, for auctions with a given number of test and control bidders, how mean revenues per auction vary with the number of test bidders. Blue dots represent mean revenues per auction, and red dots represent 95% confidence intervals around the means. Auction revenue is winsorised at the 99.9th percentile. Users who opt out of emails may participate in auctions, but are not included in either the test or control groups.

numbers of bidders and to control for the number of bidders that have opted out of emails and thus neither in test nor control. In some specifications, we also include item category fixed effects (e.g. jewelry and watches, sporting goods, musical instruments and gear). Table IV presents the results.

The regressions indicate that the email succeeded in increasing the number of bids. However, in contrast to the user level comparison of Table III, the increase in total revenue with the number of test bidders is statistically indistinguishable from zero. A simple back-of-the-envelope calculation shows the difference in magnitude between the user level point estimates of Table III and the auction level point estimates of Table IV. The average number of test and control bidders per auction is $1.40 + 1.39 = 2.79$, and sending emails to all of them would would increase revenue by 0.35%, according to the more favorable regression ($2.79 \times 0.127 \approx 0.35$). This suggests that the user level estimate of 0.74% is an overstatement of the true effect by a factor of over two.[10]

Table IV. Effect of Email on Auction Revenue and Number of Bids

	Revenue ($)		Number of Bids	
Number of Test Bidders (100β/mean)	-0.154	0.127	0.135	0.132
	(0.094)	(0.094)	(0.022)	(0.022)
Item Category Controls	No	Yes	No	Yes
N	10,425,390	10,425,390	10,425,390	10,425,390

Table shows regression coefficients on number of test bidders. The dependent variables, auction revenue and number of bids per auction, are winsorised at the 99.9th percentile. Estimates are expressed in percentage terms, relative to the mean of the corresponding dependent variable. All regressions include fixed effects for the total number of bidders participating in the experiment (test and control combined) and for the total number of other bidders.

[10]The fact that total bids change statistically significantly with test bidders, while revenue does not, can be explained by two factors. First, it is harder to detect a given percentage change in revenue than bids, as the standard deviation is larger relative to the mean for revenue than for bids (see Table II). Second, large changes in the number of bids need not translate into large changes in revenue, as each bid need not increase the final price by more than the bid increment. eBay's bid increment depends on the current high bid. For items with a current high bid between $25 and $100, for example, it is $1.

3.2. Further Evidence of Interference: Impact on Control Users

One implication of test-control interference is the difference between the user and auction level comparisons, as described above. Another implication is that control users should lose auctions more often, and have lower revenue, when bidding against test users. This is also supported by the data. In auctions with one test and one control bidder in which one of the two win, for example, the control is less likely to be the winner: 49.5% of the auctions are won by control, and 50.5% by test. If the email had no effect, we would expect test and control being equally likely to win, but the observed winning ratio is statistically different from 50%. A similar pattern is found for other test and control bidder combinations.[11]

These differences in winning rates correspond to differences in revenue. Table V presents estimates from regressions where the dependent variable is control revenue at the control user by auction level. The regressor of interest in the first specification is an indicator for the control user facing any test bidders in an auction, and in the second is the number of test bidders that the control user faces in an auction. We include fixed effects for the total number of bidders in test and control combined, and the total number of other bidders.

Both specifications indicate that control users spend substantially less when competing against test users. They spend about 1.3% less in auctions when competing against at least one test user, or 0.6% less for each test user they face. These declines are to be expected with test-control interference, and are the counterpart in our data to the decline in control revenue from \$99 to \$0 in the stylized example of Table I. This evidence suggests that control users' revenues being reassigned to test users is a likely contributor to the test-control interference bias.

Table V. Effect of Email on Control Users' Auction Revenue (\$)

Any Test Bidders (100β/mean)	-1.331 (0.441)	-
Number of Test Bidders (100β/mean)	-	-0.649 (0.152)
N	11,887,190	11,887,190

Table shows regression results on control user by auction level data. The dependent variable is the control users' revenue in an auction, and is is winsorised at the 99.9th percentile. Estimates are expressed in percentage terms, relative to the mean of the dependent variable. The regressor of interest is either a dummy for whether there are any test bidders in the auction, or the number of test bidders in the auction. Both regressions include fixed effects for the total number of bidders participating in the experiment (test and control combined) and for the total number of other bidders.

3.3. Limitations of Our Empirical Strategy

Our approach to dealing with test-control interference is to aggregate outcomes across units, mitigating interference bias. In our setting, this amounts to comparing auctions instead of users. Comparing outcomes across auctions means that within-auction interactions between users will not bias our comparisons, but there may be interference across auctions too. If a good seems likely to sell at a low price relative to substitutes which are being sold simultaneously, competition amongst bidders will tend to increase the price of that good. Selective entry of bidders into auctions of substitutes may thus reduce revenue dispersion, and reduce the likelihood of finding an effect of the email. Consequently, auctions may not be entirely unconnected experimental units either.

The ideal comparison would be between completely separate yet identical markets so that there is no test-control interference. One could compare units at a even higher level of aggregation, such as the item category (e.g. sporting goods vs musical instruments).

[11] See Table VI in the Online Appendix for the data on the other bidder combinations.

Buyers may not view goods in different categories as substitutes, thus reducing interference further. In practice, there is a bias-variance trade-off in defining the market scope. Estimates from category level comparisons are likely to be much less precise because aggregation i) reduces the sample size, in our case from millions to hundreds, and ii) reduces the experimentally induced variation in the fraction of test users across categories.[12]

For our setting, we therefore prefer our auction level specification: although there may be some residual bias remaining in our auction level estimates, the cost of eliminating it completely in terms of reduced power is high. Furthermore, there are several reasons why this kind of bidder arbitrage may not be a substantial concern in practice. First, bidders do not observe which of their competitors are test or control, making it harder to identify the auctions with weaker competition. Second, auctions rarely end simultaneously, complicating bidder arbitrage across auctions.[13] Third, even if the email campaign is successful, its effect is likely to be on the order of a few percent at most. It seems unlikely that bidder arbitrage will be so efficient that these small price differences will be eroded, given the evidence of behavioral biases that exist on eBay in related contexts.[14]

A separate concern is that the treatment may have dynamic effects. The email may have positive effects on the test group after the auction which triggered it, causing the auction level comparison to understate the email's effect. We check for differences in test and control activity in the seven days following each auction, for users who entered exactly one auction during the test period.[15] We find no significant differences in this period, suggesting that at least relative to the control group, the test group is not increasing its purchases in this period.

4. THE THEORY OF TEST-CONTROL INTERFERENCE

To develop intuition for how test-control interference generalizes beyond the perfectly inelastic unit supply and unit demand example of Table I, we first turn to a simple graphical model. We then confirm these intuitions formally, deriving a relationship between interference bias and supply and demand elasticities.

4.1. Graphical Intuition

The left diagram in Figure 3 depicts test demand before and after receiving a treatment (curves T_0 and T_1), and control demand (curve C). We assume individuals are assigned randomly and with equal probability to test and control, so that users have identical initial demand ($T_0 = C$). The right diagram shows aggregate demand before and after the treatment, along with the aggregate supply curve.

If we were to follow typical practice in A/B testing, we would estimate the effect of the experiment as test revenues minus control revenues, after the test has received the treatment. Before the treatment, revenue in both test and control is $p_0 q_0$. After the treatment, test demand shifts out and the market price increases to p_1, causing a decrease in control quantity. Post-treatment, control revenue is $p_1 q_{C,1}$ and test revenue is $p_1 q_{T,1}$. The difference, $p_1 q_{T,1} - p_1 q_{C,1}$, is the blue area $\mathcal{Z}$ in Figure 3. This is the naive A/B estimate.

We can also characterize the actual revenue change in terms of areas on the diagram. Control revenue increases by $\mathcal{X} - \mathcal{Y}$, and test revenue increases by $\mathcal{X} - \mathcal{Y} + \mathcal{Z}$. The total

[12] If the randomization is at the user level, the large number of items per category and the law of large numbers imply that there will be limited variation in the proportion of test users from one category to another.

[13] Successful arbitrage requires assessing whether an item is underpriced given the time remaining in the auction, a much more complicated task than simply comparing current prices of different goods.

[14] Hossain and Morgan [2006], Brown et al. [2010] and Einav et al. [2011] find that consumers respond more to increases in the price of the item net of shipping fees than to increases in shipping fees. Backus et al. [2013] find that eBay users are prone to arguably irrational incremental bidding.

[15] Table VII in the Online Appendix displays these results.

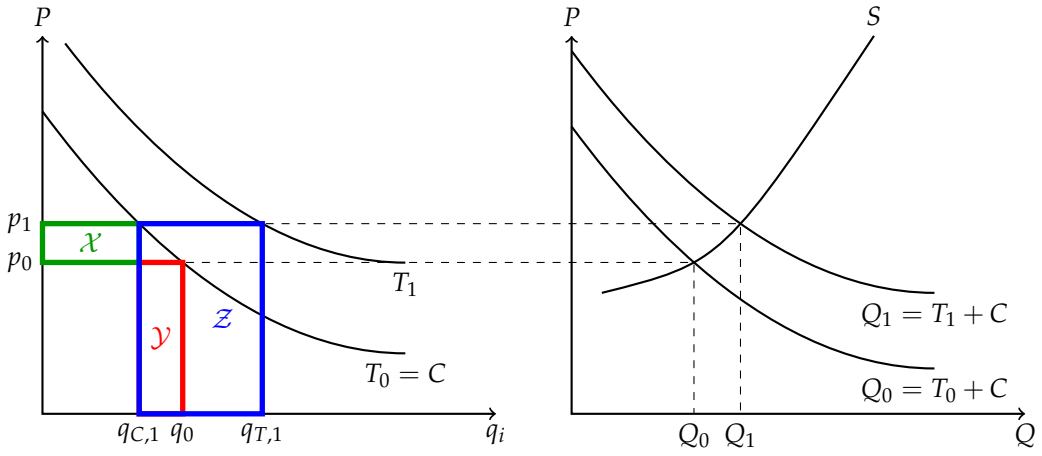

Fig. 3. Aggregate Supply Effects

increase in revenue produced by the experiment is the sum of these, $2(\mathcal{X} - \mathcal{Y}) + \mathcal{Z}$.[16] The interference bias is the biased user level difference minus this amount, that is, $\mathcal{Z} - (2(\mathcal{X} - \mathcal{Y}) + \mathcal{Z}) = 2(\mathcal{Y} - \mathcal{X})$.

There are two different, countervailing interference biases corresponding to the areas $\mathcal{X}$ and $\mathcal{Y}$. First, the usual test-control comparison ignores the fact that the experiment causes prices to increase, raising revenue from both test and control users that continue to buy (area $\mathcal{X}$). This creates a tendency to understate the market-wide revenue gains, contributing a downward bias. Second, the approach ignores the fact that the higher prices caused by the experiment result in decreases in quantity purchased (area $\mathcal{Y}$). This contributes an upward bias. This area $\mathcal{Y}$ is the revenue "reassignment" effect referred to in previous sections: without the experiment, the control quantity would be q_0, not $q_{c,1}$.

The sign of the bias depends on the the relative sizes of the areas $\mathcal{X}$ and $\mathcal{Y}$, which is determined by demand elasticity before the experiment. If demand is elastic, then increasing the price from p_0 to p_1 reduces revenue, so $\mathcal{Y} - \mathcal{X} > 0$ and the bias is positive.[17,18] If demand is inelastic, $\mathcal{Y} - \mathcal{X} < 0$ and the bias is negative. Given the degree of competition in e-commerce, one might expect price sensitive consumers and elastic demand. This is consistent with our empirical findings above: the difference between the user and auction level comparisons suggests that the bias is indeed positive.

While the sign of the bias depends on the elasticity of demand, its magnitude depends on the magnitude of the price change, and hence on the elasticity of supply. As supply becomes more elastic, the size of the price change caused by the increase in aggregate demand shrinks, as do the areas $\mathcal{X}$ and $\mathcal{Y}$.[19] With perfectly elastic supply, the increase in aggregate demand generated by the experiment does not change prices at all. Control revenue is the same before and after the experiment; the areas $\mathcal{X}$ and $\mathcal{Y}$ disappear, and the

[16] In terms of the right-hand side diagram, this is also equal to $(p_1 \times Q_1) - (p_0 \times Q_0)$.

[17] More precisely if $C(p)$ denotes the (continuously differentiable) pre-treatment control group demand curve and $\varepsilon(p)$ is its elasticity of demand, the fundamental theorem of calculus implies that $p_1 \times q_{C,1} < p_0 \times q_0 \Leftrightarrow \int_{p_0}^{p_1} C(p)(1+\varepsilon(p))\,dp < 0$. A sufficient condition for revenue to fall from p_0 to p_1 is that $\varepsilon(p) < -1$ for $p \in [p_0, p_1]$.

[18] Our example in Table I is an extreme instance of this case, as control demand is perfectly elastic. The price increase from the email makes control demand drop to zero, so that the area $\mathcal{X}$ is zero and the bias is positive.

[19] Blake et al. [2013] limit their analysis to eBay fixed price (non-auction) transactions when evaluating a marketing experiment for this reason. Fixed price goods tend to be more homogeneous commodities and are likely to be more elastically supplied.

test-control comparison gives the correct answer. Intuitively, with perfectly elastic supply, the experimentally generated increase in test demand does not affect the supply available to the control group at all, so there is no test-control interference. The more inelastic aggregate supply is, the greater the price increase a given shift in demand generates, and the larger the areas $\mathcal{X}$ and $\mathcal{Y}$ are.

Finally, note that inferring what would happen if both test and control groups received the treatment requires projecting out-of-sample. If the control received the treatment too, the aggregate demand curve would increase beyond Q_1 and price would increase beyond p_1. By how much prices would increase beyond p_1 is not something the A/B test can answer without making assumptions on the shape of the aggregate supply curve, as it gives no information on the slope of supply after Q_1. Inferring the effect of treating the whole population therefore involves another source of bias, which is introduced by extrapolating out-of-sample.

4.2. Expressing Relative Bias in Terms of Elasticities

We derive an approximation to the *relative bias*, or the ratio of the interference bias to the naive treatment estimate. Let there be a unit mass of consumers with pre-treatment aggregate demand function $D(p)$, and let $S(p)$ denote the aggregate supply function. A fraction t of consumers is randomly selected to be in the test group. Before the treatment the test group's demand is $tD(p)$ and the control group's demand is $(1-t)D(p)$. Suppose the treatment increases test demand multiplicatively, so that after the treatment, the test demand function is $(1+\alpha)tD(p)$ and the aggregate demand function is $(1+\alpha)tD(p) + (1-t)D(p) = (1+\alpha t)D(p)$. As above, let p_0 be the equilibrium price before the treatment, and p_1 be the equilibrium price afterwards.

The naive estimate of the effect of the experiment on revenue is $\alpha t D(p_1)p_1$, as on average each test user buys $\alpha D(p_1)$ more units than control users do at the post-treatment price of p_1, and test users are a fraction t of the population. The true effect of the experiment is $(1+\alpha t)D(p_1)p_1 - D(p_0)p_0$. The bias is $\alpha t D(p_1)p_1 - (1+\alpha t)D(p_1)p_1 + D(p_0)p_0 = D(p_0)p_0 - D(p_1)p_1$.

We define the relative bias to be the ratio of the bias to the naive user level estimate, that is, $(D(p_0)p_0 - D(p_1)p_1)(\alpha t D(p_1)p_1)^{-1}$. Let $\varepsilon_S(p)$ and $\varepsilon_D(p)$ denote the price elasticities of $S(p)$ and $D(p)$. The following theorem approximates the relative bias in terms of these supply and demand elasticities.

THEOREM 4.1. *If aggregate supply $S(p)$ is strictly increasing, pre-treatment aggregate demand $D(p)$ is strictly decreasing, and both are continuously differentiable, then the relative bias* $\frac{D(p_0)p_0 - D(p_1)p_1}{\alpha t D(p_1)p_1} = -\frac{1+\varepsilon_D(p_1)}{\varepsilon_S(p_1)-\varepsilon_D(p_1)} + O(\alpha)$.

PROOF. See the Online Appendix.

This expression for relative bias confirms the graphical analysis: the relative bias is positive when demand is elastic and negative when demand is inelastic; it is zero when supply is perfectly elastic and increases in magnitude when supply becomes more inelastic. The absolute bias, $D(p_0)p_0 - D(p_1)p_1$, rises with the effectiveness of the treatment (α) and the fraction of people treated (t), because both factors increase the post-treatment price p_1. However, the naive estimate rises proportionally so the relative bias is not affected by these parameters. Figure 4 shows our approximation of the relative bias for various supply and demand elasticities. In practice, most experimentation is likely to be performed in competitive marketplaces with elastic demand so the relative bias from the naive estimate may be large.

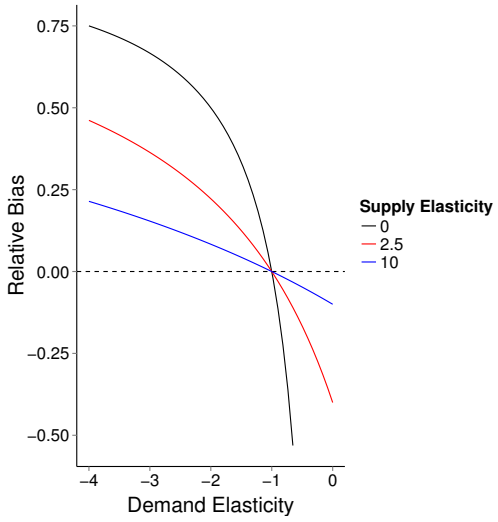

Fig. 4. Relative Bias and Elasticities

5. CONCLUSION

While the email experiment we study provides a particularly stark example of test-control interference, this kind of interference is likely to affect experimentation in marketplaces quite generally. It introduces a bias in the naive test-control comparison which will be especially severe when supply is inelastic. If demand is elastic the bias will be positive, so that the naive comparison overstates the treatment's benefit. Mitigating this bias is not a trivial task, but a better strategy may be to compare units at a higher level of aggregation (e.g. auctions instead of individuals), across which there is less pronounced interference. Developing other strategies for dealing with this bias is likely to be valuable for future analysis of experiments in markets.

ACKNOWLEDGMENTS

We thank Steven Tadelis, and the eBay traffic sciences team for their guidance and support. We also thank the program committee and our reviewers for their very thoughtful comments.

REFERENCES

BACKUS, M., BLAKE, T., MASTEROV, D., AND TADELIS, S. 2013. Is sniping a problem for online auction markets?

BASSLER, D., MONTORI, V. M., BRIEL, M., GLASZIOU, P., AND GUYATT, G. 2008. Early stopping of randomized clinical trials for overt efficacy is problematic. *Journal of clinical epidemiology 61,* 3, 241–246.

BENJAMINI, Y. AND HOCHBERG, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

BLAKE, T., NOSKO, C., AND TADELIS, S. 2013. Consumer heterogeneity and paid search effectiveness: A large scale field experiment. *NBER Working Paper*, 1–26.

BROWN, J., HOSSAIN, T., AND MORGAN, J. 2010. Shrouded attributes and information suppression: Evidence from the field. *The Quarterly Journal of Economics 125,* 2, 859–876.

COX, D. R. 1958. *Planning of experiments*. Wiley, New York.

CROOK, T., FRASCA, B., KOHAVI, R., AND LONGBOTHAM, R. 2009. Seven pitfalls to

avoid when running controlled experiments on the web. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1105–1114.
DAVID, O. AND KEMPTON, R. A. 1996. Designs for interference. *Biometrics*, 597–606.
DUNN, O. J. 1961. Multiple comparisons among means. *Journal of the American Statistical Association 56*, 293, 52–64.
EINAV, L., KUCHLER, T., LEVIN, J. D., AND SUNDARESAN, N. 2011. Learning from seller experiments in online markets. Tech. rep., National Bureau of Economic Research.
FISHER, R. A. 1935. The design of experiments.
FRADKIN, A. 2013. Search frictions and the design of online marketplaces.
HOSSAIN, T. AND MORGAN, J. 2006. ... plus shipping and handling: Revenue (non) equivalence in field experiments on eBay. *Advances in Economic Analysis & Policy 5*, 2.
KERSHNER, R. P. AND FEDERER, W. T. 1981. Two-treatment crossover designs for estimating a variety of effects. *Journal of the American Statistical Association 76*, 375, 612–619.
KOHAVI, R., LONGBOTHAM, R., SOMMERFIELD, D., AND HENNE, R. M. 2009. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery 18*, 1, 140–181.
LEWIS, R. A. AND RAO, J. M. 2013. On the near impossibility of measuring the returns to advertising.
LUCKING-REILEY, D. 1999. Using field experiments to test equivalence between auction formats: Magic on the internet. *American Economic Review*, 1063–1080.
MONTGOMERY, D. C. 1984. *Design and analysis of experiments*. Vol. 7.
REILEY, D. H. 2006. Field experiments on the effects of reserve prices in auctions: More magic on the internet. *The RAND Journal of Economics 37*, 1, 195–211.
ROSENBAUM, P. R. 2007. Interference between units in randomized experiments. *Journal of the American Statistical Association 102*, 477.
RUBIN, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*, 5, 688.
RUBIN, D. B. 1986. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association 81*, 396, 961–962.
SIMMONS, J. P., NELSON, L. D., AND SIMONSOHN, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science 22*, 11, 1359–1366.
SINCLAIR, B., MCCONNELL, M., AND GREEN, D. P. 2012. Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science 56*, 4, 1055–1069.

Online Appendix to:
Why Marketplace Experimentation Is Harder than it Seems: The Role of Test-Control Interference

THOMAS BLAKE, eBay Research Labs
DOMINIC COEY, eBay Research Labs

PROOF OF THEOREM 4.1. Define $\phi(p, \alpha) = S(p) - (1 + \alpha t)D(p)$. By the implicit function theorem, there is a unique function $p(\alpha)$ defined in a neighborhood $\mathcal{N}$ of $\alpha = 0$ such that for all $\alpha \in \mathcal{N}$, $\phi(p(\alpha), \alpha) = 0$, and this function has derivative $p'(\alpha) = \frac{tD(p(\alpha))}{S'(p(\alpha)) - D'(p(\alpha))}$.

Define $R(\alpha) = D(p(\alpha))p(\alpha)$, and recall that by definition $p_0 = p(0)$ and $p_1 = p(\alpha)$. A first-order Taylor expansion of R around α gives

$$R(0) - R(\alpha) = -\alpha \left(D'(p_1)p_1 + D(p_1)\right) \cdot \frac{tD(p_1)}{S'(p_1) - D'(p_1)} + O(\alpha^2).$$

Our expression for the relative bias is therefore

$$\begin{aligned}
\frac{D(p_0)p_0 - D(p_1)p_1}{\alpha t D(p_1)p_1} &= \frac{R(0) - R(\alpha)}{\alpha t D(p_1)p_1} \\
&= -\frac{(D'(p_1)p_1 + D(p_1))}{p_1} \cdot \frac{1}{S'(p_1) - D'(p_1)} + O(\alpha) \\
&= -\frac{\frac{1}{D(p_1)}\left(D'(p_1)p_1 + D(p_1)\right)}{\frac{p_1}{D(p_1)}\left(S'(p_1) - D'(p_1)\right)} + O(\alpha) \\
&= -\frac{1 + \varepsilon_D(p_1)}{\varepsilon_S(p_1) - \varepsilon_D(p_1)} + O(\alpha). \quad \square
\end{aligned}$$

Table VI. Fraction of Auctions Won by Test, Conditional on Test or Control Winning

Bidders	N	Actual	If No Effect
1 T, 1 C	852,258	50.50 (0.05)	50.00
1 T, 2 C	417,844	33.67 (0.07)	33.33
2 T, 1 C	422,114	66.93 (0.07)	66.66
1 T, 3 C	187,027	25.23 (0.10)	25.00
2 T, 2 C	280,426	50.15 (0.09)	50.00
3 T, 1 C	188,656	75.20 (0.10)	75.00

"Actual" column shows actual fraction of auctions won by test, conditional on test or control winning, for different configurations of test and control entry. Standard errors are in parentheses. "If No Effect" column shows what the expected fraction of auctions won by test would be, if the email has no effect.

Table VII. Treatment Effects on Subsequent Purchase Activity

	Total Revenue	Item Count
Test Group (100β/mean)	0.599 (0.755)	0.195 (0.340)
Observations	1,788,281	1,788,281

Table shows regression results on user level data. To focus on the follow-on effects of the email, we limit the sample to users who received exactly one email from an auction they entered during the experiment period. The dependent variables are "total revenue" and "item count". "Total revenue" is the total value of user purchases made in the seven days after the end of the auction for which they received the email. "Item count" is the number of other auctions entered by the user (for which they did not receive an email, because their first bid was after the email was sent) plus the number of fixed price items bought over this time. Estimates are expressed in percentage terms, relative to the mean of the dependent variable. Total revenue is winsorised at the 99.9th percentile.

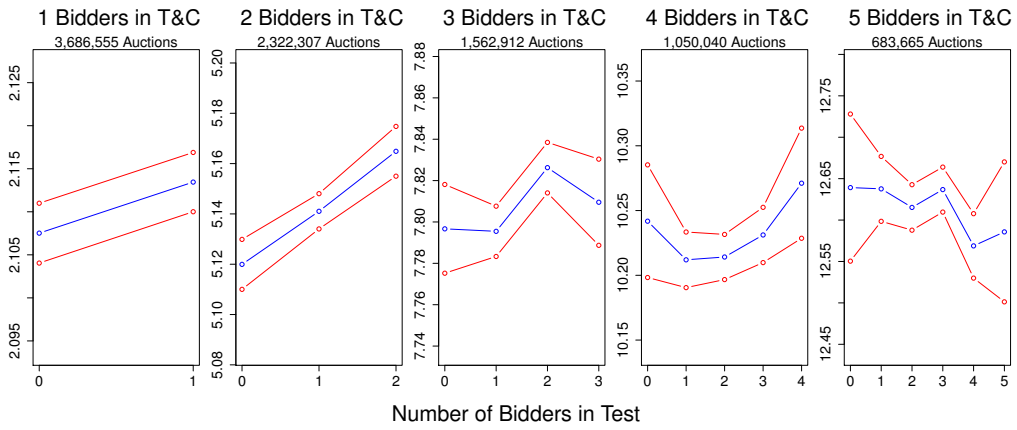

Fig. 5. The Effect of Emails on Total Bids

Plots show, for auctions with a given number of test and control bidders, how the total number of bids varies with the number of test bidders. Blue dots represent mean total bids per auction, and red dots represent 95% confidence intervals around the means.

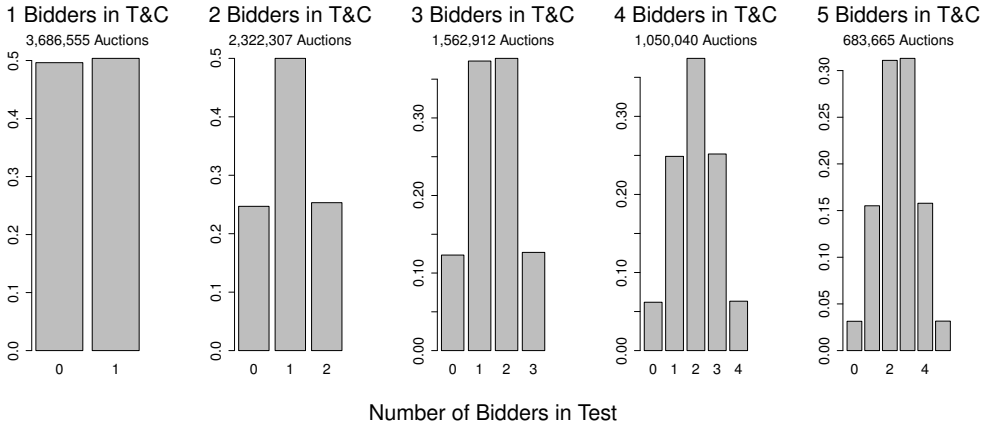

Fig. 6. Histogram by Auction Size

Plots show, for auctions with a given number of test and control bidders, the distribution of the number of test bidders per auction.

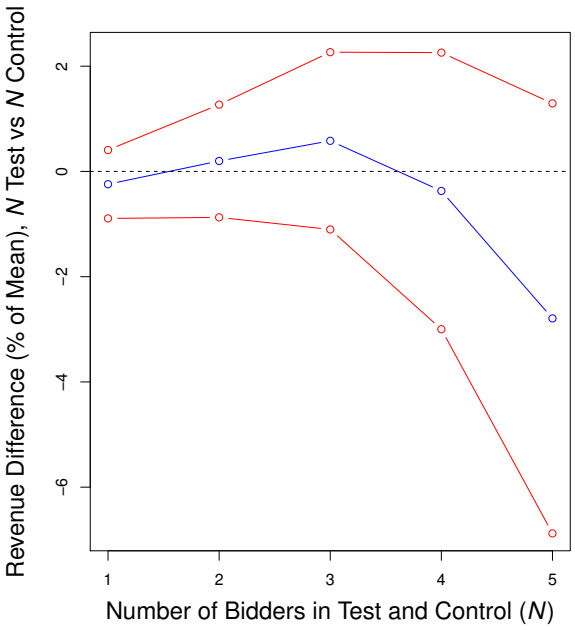

Fig. 7. The Effect of Emails on Total Revenue, N Test vs N Control

Figure shows differences in revenue between auctions with N test and 0 control users, and 0 test and N control users, for $N = 1 \ldots 5$. Blue dots represent mean revenue differences per auction, expressed as a percentage of mean revenue in auctions with the corresponding number of test and control bidders, and red dots represent 95% confidence intervals around the means.