SUPPLEMENT TO "CONSUMER HETEROGENEITY AND PAID SEARCH EFFECTIVENESS: A LARGE-SCALE FIELD EXPERIMENT": APPENDIX

(*Econometrica*, Vol. 83, No. 1, January 2015, 155–174)

BY THOMAS BLAKE, CHRIS NOSKO, AND STEVEN TADELIS

BRAND KEYWORD CLICK MAGNITUDE ESTIMATION

TO QUANTIFY THE SUBSTITUTION between natural and organic search for brand terms, we first regress the log of total daily clicks from MSN to eBay on an indicator for whether days were in the period with ads turned off. Results are shown in Table A.I. Click volume was 5.6 percent lower in the period after advertising was suspended. We then use data on eBay's clicks from Google as a control for seasonal factors because during the test period on MSN, eBay continued to purchase brand keyword advertising on Google. With these data, we calculate the change in total click traffic in the presence of brand keyword advertising. In the difference-in-differences approach, we add observations of daily traffic from Google and Yahoo! and include in the specification search engine dummies and trends.[1] The variable of interest is thus the interaction between a dummy for the MSN platform and a dummy for treatment (ad off) period. Column (2) of Table A.I shows a much smaller effect once the seasonality is accounted for. In fact, only 0.529 percent of the click traffic is lost, so 99.5 percent is retained. Notice that this is a lower bound of retention because some of the 0.5 percent of traffic that no longer comes through Google may be switching to non-Google traffic (e.g., typing "ebay.com" into the browser).

NON-BRAND KEYWORD SALES MAGNITUDE ESTIMATION

To determine the size of the effect of paid search on sales, we estimate difference-in-differences and generalized fixed models as follows:

$$\ln(Sales_{it}) = \beta_1 \times AdsOn_{it} + \beta_2 \times Post_t + \beta_3 \times Group_i + \varepsilon_{it}, \tag{A.1}$$

$$\ln(Sales_{it}) = \beta_1 \times AdsOn_{it} + \delta_t + \gamma_i + \varepsilon_{it}. \tag{A.2}$$

In this specification, i indexes the DMA, t indexes the day, $Post_t$ is an indicator for whether the test was running, $Group_i$ is an indicator equal to 1 if region i kept search spending on, and $AdsOn_{it}$ is the interaction of the two indicators. In the second specification, the base indicators are subsumed by day and DMA fixed effects. The β_1 coefficient on the interaction term is then the percentage

[1]The estimates presented include date fixed effects and platform specific trends, but the results are very similar without these controls.

 DOI: 10.3982/ECTA12423

TABLE A.I

QUANTIFYING BRAND KEYWORD SUBSTITUTION[a]

	MSN		Google
	(1) Log Clicks	(2) Log Clicks	(3) Log Clicks
Period	−0.0560***		−0.0321*
	(0.00861)		(0.0124)
Interaction		−0.00529	
		(0.0177)	
Google		5.088	
		(10.06)	
Yahoo!		1.375	
		(5.660)	
Constant	12.82***	11.33*	14.34***
	(0.00583)	(5.664)	(0.00630)
Date FE		Yes	
Platform Trends		Yes	
N	118	180	120

[a]Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. This table shows regression analyses of the data in Figure 2. Column (1) shows estimates of a regression of the log of daily click counts on an indicator for whether the day is after the brand keyword suspension. Column (2) adds daily data from Google and Yahoo! and shows the estimates of a difference-in-differences regression. Column (3) mimics Column (1) for the Google suspension and resumption of brand keyword advertising.

effect on sales because the $Sales_{it}$ is the log of total sales in region i on day t. We restrict attention to sales from fixed-price transactions because auctions may pool users from both test and control DMAs, which in turn would attenuate the effect of ads on sales.[2] We control for inter-DMA variation with DMA clustered standard errors and DMA fixed effects.

Columns (1) and (2) in Table A.II correspond to Equations (A.1) and (A.2), respectively, where an observation is at the daily DMA level, resulting in 23,730 observations. Columns (3) and (4) correspond to Equations (A.1) and (A.2), respectively, where an observation is aggregated over days at the DMA level for the pre and post periods separately, resulting in 420 observations. All regression results confirm the very small and statistically insignificant effect of paid search ads.

PRODUCT RESPONSE HETEROGENEITY

A consumer's susceptibility to Internet search ads depends on how well informed they are about where such products are available. Given that the avail-

[2]The results throughout are quantitatively similar even if we include auction transactions.

TABLE A.II
DIFF-IN-DIFF REGRESSION ESTIMATES[a]

	Daily		Totaled	
	(1) Log Sales	(2) Log Sales	(3) Log Sales	(4) Log Sales
Interaction	0.00659	0.00659	0.00578	0.00578
	(0.00553)	(0.00555)	(0.00572)	(0.00572)
Experiment Period	−0.0460***		0.150***	0.150***
	(0.00453)		(0.00459)	(0.00459)
Search Group	−0.0141		−0.0119	
	(0.168)		(0.168)	
DMA Fixed Effects		Yes		Yes
Date Fixed Effects		Yes		
N	23,730	23,730	420	420

[a]Standard errors, clustered on the DMA, in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table shows regression analyses of the data in Figure 3(b). Columns (1) and (2) present DMA by day level regressions, where Columns (3) and (4) aggregate to the DMA by pre-post level. The interaction term is the effect of the total spending on sales.

ability of products varies widely, the effectiveness of paid search may vary by product type. As a large e-commerce platform, eBay's paid search advertising campaigns present an opportunity to test the returns to advertising across product categories which vary in competitiveness, market thickness, and general desirability. To our surprise, different product attributes did not offer any significant variation in paid search effectiveness.

As in Section 3.2, we decompose the response by interacting the treatment indicator with dummies for sub-groupings of revenue using the category of sales. We found no systematic relationship between returns and category. The estimates center around zero and are generally not statistically significant. At the highest level, only one category is significant, but with 38 coefficients, at least one will be significant by chance.

We explored multiple layers of categorization, ranging from the broadest groupings of hundreds of categories. The extensive inventory eBay offers suggests that some categories would generate returns because customers would be unaware of their availability on eBay. However, we have looked for differential responses in a total of 378 granular product categories and found no consistent pattern of response. Less than 5 percent of categories are statistically significant at the 5 percent confidence level. Moreover, in an examination of the estimates at finer levels of categorization, we found no connection between ordinal ranking of treatment effect product features like sales volume or availability. It is thus evident that for a well-known company like eBay, product attributes are less important in search advertising than user intent and, more importantly, user information.

CONTROLLED BRAND KEYWORD EXPERIMENTS

In January 2013, we conducted a follow-up test of the brand term paid search, specifically keyword eBay, using geographic variation. Google offers geographic specific advertising across Germany's 16 states. So we selected a random half of the country, 8 states, where brand keyword ads were suspended. This test design preserves a randomly selected control group which is absent from the simple pre-post analysis shown in Section 3.

As was predicted by the earlier tests, there was no measurable effect on revenues. The sample size for this analysis is smaller because there are fewer separable geographical areas and the experiment window is shorter. Figure A.1 shows the log difference between means sales per day in the on and off states. The treatment group is 5 percent smaller, on average, than the control because there are few states, so any random division of states generates a baseline difference. The plot shows that there is no change in the difference once the experiment begins. The plot also shows the large variation in daily differences of the means, which suggests that detecting a signal in the noise would be very difficult.

We perform a difference-in-differences estimation of the effect of brand advertising on sales and find no positive effect. Table A.III presents the results from three specifications: the baseline model from Section 4.1 in Column (1), the same specification with the (less noisy) subset of data beginning January 1, 2013 in Column (2), and the smaller subset with state specific linear time trends in Column (3). All results are small, statistically insignificant, and negative.

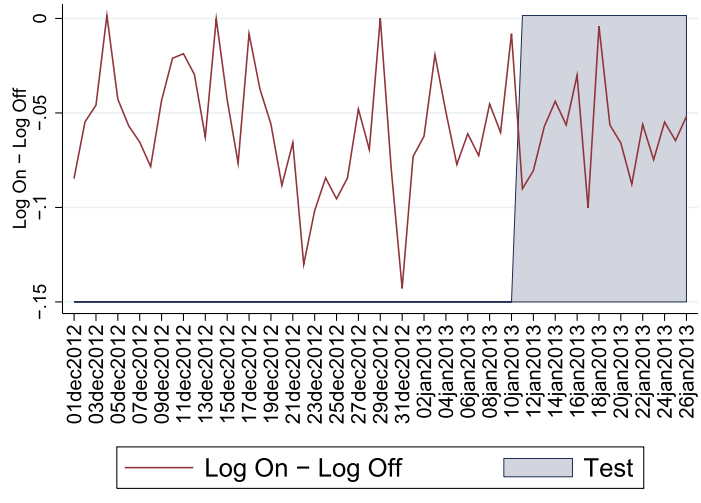

FIGURE A.1.—Brand test Europe: Difference in total sales. This figure plots the difference between on and off regions in Germany before and after the experimental shut off of brand advertising.

TABLE A.III

BRAND TEST EUROPE: DIFFERENCE-IN-DIFFERENCES ESTIMATES[a]

	(1) Log Sales	(2) Log Sales	(3) Log Sales
Interaction	−0.00143	−0.00422	−0.000937
	(0.0104)	(0.0132)	(0.0140)
State Fixed Effects	Yes	Yes	Yes
Date Fixed Effects	Yes	Yes	Yes
Post Jan 1		Yes	Yes
State Trends			Yes
N	912	416	416

[a]Standard errors, clustered on the state, in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. This table presents the regression analysis of Figure A.1. Column (1) is a difference-in-differences estimation with region (state) and date fixed effects. Columns (2) and (3) add additional time controls and trends.

Narrowing the window and controlling for state trends reduces the magnitude of the estimate, which is consistent with a zero result. This test is noisier than the U.S. test because there are substantially fewer (16) geographic regions available for targeting. This makes the confidence intervals larger. We also lack public estimates of spending for Germany and are therefore unable to derive a confidence interval around the ROI which is likely to be large anyway due to the smaller total spending levels for branded advertising. The negative point estimates support the findings of the U.S. brand spending changes and lead us to conclude that there is no measurable or meaningful short run return to brand spending.

RANDOMIZATION PROCEDURE DETAIL

The treatment assignment used a stratification procedure to ensure common historical trends between treatment and control. This means that the treatment dummy is not a simple random variable, but instead lends itself to a difference-in-differences estimation. The test regions, or DMAs, were chosen in two steps. First, 133 of the 210 U.S. regions were selected to be candidates for treatment purely at random. Of these, only about half were allotted to be treated (advertising turned off). To minimize historical variation between test and control, groups of 68 were drawn at random and then the historical weekly serial correlation was computed. Several draws with very low historical correlation were discarded before the current draw of 68 in one group and 65 in another. Which group was actually turned off was decided by a coin flip. The 68 regions were then turned off at an arbitrary date (based largely on engineering availability). This procedure lends itself perfectly to a difference-in-differences estimation where the core underlying assumption is common trends.

CANDIDACY IV ESTIMATION

The assignment of DMAs into treatment cells for the non-brand keyword experiment was stratified on historical trends. This stratification lacks the clarity of a total random assignment, but the methodology admits an alternative approach that leverages the completely random assignment to the set of DMAs eligible for testing. We use the assignment to the *candidate* group as an instrument for whether or not a DMA was assigned to the treatment group. We collapse the data to the DMA level and use two stage least squares to estimate the effect of treatment assignment and advertising spending on revenue. We include pre-period sales in both stages to control for variations in DMA size. The first and second stages are as follows:

$$\ln(Sales_i) = \beta_0 + \beta_1 \times AdsOn_i + \beta_2 \times \ln(PreSales)_i + \varepsilon_i, \tag{A.3}$$

$$AdsOn_i = \alpha_0 + \alpha_1 \times Candidate_i + \mu_i. \tag{A.4}$$

The estimates are shown in Table A.IV. The coefficients on both extensive (*AdsOn*) and intensive (spending level) are smaller than those in Table A.II and Table I, respectively. If the stratification assignment introduces a bias into the treatment effect, it is a positive bias which makes our primary estimates an upper bound on the true effect of advertising on paid search. The standard errors

TABLE A.IV

CANDIDATE DMA INSTRUMENT[a]

	(1) ln(Test Period Sales)	(2) ln(Test Period Revenue)
Ads On	0.00207	
	(0.0108)	
ln(Test Period Spend)		0.000795
		(0.00350)
ln(Pre Period Sales)	1.007***	
	(0.00226)	
ln(Pre Period Revenue)		0.997***
		(0.0113)
ln(Pre Period Spend)		0.00877
		(0.0109)
Constant	0.0436	0.102**
	(0.0354)	(0.0520)
Observations	210	210

[a]Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table presents estimates from a two stage least squares estimation where an indicator for whether the DMA was a 'candidate' for ad suspension is used as an excluded instrument for whether ads were left on (in Column (1)) and test period spending (Column (2)).

in this IV approach are larger than primary specifications, making this method less precise. The loss of precision in the IV stems from the reduction in observations from 420 in Table A.II to 210 because the difference-in-differences approach uses the exogenous timing of the treatment. Moreover, the stratification was in fact designed to reduce intertemporal variance across treatment cells.

CANDIDACY DIFFERENCE-IN-DIFFERENCES

To further check the randomization procedure, we estimated the difference-in-differences using the candidacy indicator as the treatment dummy. This would estimate the diluted effect on all DMAs that were considered candidates for testing. Column (1) of Table A.V shows the results. The negative coefficient here is expected since 'candidate' DMAs were candidates for turning ads off. Thus, the 0.25 percent is just under half the magnitude of the main estimates we present.

To check the randomization, we repeated this estimation on the subsample of DMAs that were not selected into treatment. These DMAs were all controls in the main estimation and ad spending was on throughout the experiment. The coefficient therefore represents the change in the candidate control regions over the non-candidate control regions during the test. Column (2) of Table A.V presents the results, a small positive coefficient which is statistically

TABLE A.V

CANDIDATE DMA DIFF-IN-DIFF[a]

	(1) Log Sales	(2) Log Sales
Interaction	−0.00247	0.00124
	(0.00526)	(0.00653)
Candidate for off DMA	0.167***	1.194***
	(0.00288)	(0.00358)
Experiment Period	−0.199***	−0.195***
	(0.0117)	(0.0150)
Constant	11.87***	11.87***
	(0.00738)	(0.00924)
Date Fixed Effects	Yes	Yes
DMA Fixed Effects	Yes	Yes
Ads On Only DMAs		Yes
N	23,730	16,046

[a]Standard errors, clustered on the DMA, in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table presents estimates from a difference-in-differences analysis just as in Table A.II where the indicator for whether spending was suspended is replaced by an indicator for whether the region was a 'candidate' for suspension.

TABLE A.VI
ROI IN LEVELS[a]

	(1) Revenue ($)
Cost ($)	0.199
	(0.161)
DMA Fixed Effects	Yes
Date Fixed Effects	Yes
N	23,730

[a]Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. This table presents estimates similar to Table I, Column (4) where revenue and spending levels are used in place of logged values.

zero. We conclude that the stratification procedure did not create any evidence of a biased treatment assignment.

ROI RESULTS IN LEVELS

The preferred specification reports regressions in logged dependent variable because the dependent variable is very right skewed (some DMAs are very large and some days have large positive shocks). Given the large variance across days and DMAs, the average is not particularly meaningful. The IV approach of Equations (1) and (2) permits a more straightforward estimation of ROI using level regressions. Table A.VI presents the result of an estimation of Equations (1) and (2) where $\ln(Spend)$ and $\ln(Rev)$ are replaced with spend and revenue, respectively, in dollars. The coefficient can be interpreted as the dollar increase in revenue for every dollar spent. A coefficient of 0.199 implies a ROI of -80 percent, comparable but slightly more negative than the primary results of -63 percent. The confidence interval of this estimate excludes the break-even point of 1. The levels estimate is qualitatively similar to the log results and so we present the more conservative estimation as our preferred specification.

ROI CALCULATIONS

Recall that ROI is defined as

$$ROI = \frac{R_1 - R_0}{S_1 - S_0} - 1 \equiv \frac{\Delta R}{\Delta S} - 1. \tag{A.5}$$

Let R_i be the revenue in DMA i and let $D_i = 1$ if DMA i was treated (paid search off) and $D_i = 0$ if it was not (paid search on). The basic difference-in-difference regression we ran is

$$\ln(R_{it}) = \beta_1 D_i + \delta_t + \gamma_i + \varepsilon_{it},$$

where δ_t and γ_i are time and DMA fixed effects. Using the natural logarithm $\ln R_{it}$ implies that, for small differences in $R_1 - R_0$, the coefficient β_1 in the regression is approximately the percent change in revenue for the change in the spend level that results from the experimental treatment. This means that, for two revenue levels R_1 and R_0, we can write

$$\beta_1 \approx \frac{R_1 - R_0}{R_0}$$

or

$$R_0 \approx \frac{R_1}{1 + \beta_1}. \tag{A.6}$$

Because the spend in the "off" DMAs is $S_0 = 0$ (or close to it) and in the "on" DMAs is some S_1, then, using (A.6) and (A.5), we can derive the approximate ROI as

$$ROI = \frac{R_1 - R_0}{S_1 - S_0} - 1 \approx \frac{R_1 - \dfrac{R_1}{1 + \beta_1}}{S_1} - 1 = \frac{\beta_1}{(1 + \beta_1)} \frac{R_1}{S_1} - 1. \tag{A.7}$$

Thus, Equation (A.7) gives a well-defined and financially correct estimate of the ROI based on the difference-in-difference estimate of the experimental results when $S_0 = 0$.

Unlike the difference-in-differences estimates, the OLS and IV estimates were derived from the regression

$$\ln(R_{it}) = \alpha_1 \ln(S_{it}) + \varepsilon_{it},$$

where the first stage of the IV estimation used the regression

$$\ln(S_{it}) = \tilde{\alpha}_1 [AdsOn_{it}] + \tilde{\alpha}_2 [Post_t] + \tilde{\alpha}_3 [Group_i] + \varepsilon_{it}.$$

From these, we find $\alpha_1 = \frac{\Delta \ln(Sales)}{\Delta \ln(Spend)}$ and $\tilde{\alpha}_1 = \Delta \ln(Spend)$ for which the approximation to a percentage change is poor since the change in spend was large. Therefore, to make use of the log-log regression coefficients, it is possible to translate them into a reduced form effect because

$$\alpha_1 * \tilde{\alpha}_1 = \frac{\Delta \ln(Sales)}{\Delta \ln(Spend)} * \Delta \ln(Spend) = \Delta \ln(Sales) = \beta_1,$$

which we can substitute into (A.7). Thus, for the estimated α_1 of the OLS estimates, we can use $\tilde{\alpha}_1$ to derive a comparable β_1 which can be used to compute an ROI that is comparable across all specifications.

eBay Research Labs, 2065 Hamilton Ave., San Jose, CA 95125, U.S.A.; thblake@ebay.com,

University of Chicago, 5807 South Woodlawn Ave., Chicago, IL 60637, U.S.A. and eBay Research Labs; cnosko@chicagobooth.edu,

and

UC Berkeley, 2220 Piedmont Ave., Berkeley, CA 94720-1900, U.S.A., NBER, and eBay Research Labs; stadelis@berkeley.edu.

Manuscript received April, 2014; final revision received August, 2014.