

Messaging and the Mandate: The Impact of Consumer Experience on Health Insurance Enrollment Through Exchanges

By NATALIE COX, BENJAMIN HANDEL, JONATHAN KOLSTAD, AND NEALE MAHONEY*

At the heart of the Affordable Care Act (ACA) are efforts to expand insurance coverage. A key element of this effort are “insurance exchanges,” which are designed to give consumers access to a variety of insurance options at reasonable prices and to administer price subsidies when applicable. Doing so requires these exchanges to serve a number of functions: making a market, determining the number of options to offer, determining the degree to which products and prices can be differentiated, and providing decision and enrollment support (e.g., Ericson and Starc, 2013).

Accomplishing all of these goals has not been easy. The Federal website, HealthCare.gov, was plagued by glitches during its first months of operation and a number of state exchanges were hampered by similar challenges (e.g. Oregon). At the same time, a number of private exchanges have developed or expanded to serve these populations (e.g. eHealth). Whether and how competition between public and private exchanges impacts welfare is a critical policy question. In this paper we focus on a specific element of the distinction between public and private exchanges: the ability of private exchanges to run large-scale experiments and use the findings to target messaging to consumers.

The use of large-scale experiments, particularly in online settings, has become widespread in the private sector (e.g., Brown, Chui and Manyika, 2011). One goal of such experimentation is to understand and enhance the experience of customers. An important distinction between web-based experimentation and more traditional experiments in marketing is the ability to measure heterogeneous responses (given the scale of the randomization) and subsequently provide specific consumers with different messaging (e.g., Goldfarb, 2014). Outside of health insurance choice, this heterogeneity has been found to be substantial. For example, in a study of the impact of marketing messages, Lewis and Reily (2014) find that the over-65 population accounted for only 5 percent of the sample but generated 40 percent of the overall impact. Blake, Nosko and Tadelis (forthcoming) demonstrate that paid search, one method of providing targeted information to web consumers, has little impact on average but substantial returns for new and infrequent users.

A number of factors make health insurance exchanges ripe for experimentation and targeting. The population of interest in health insurance exchanges—previously uninsured shoppers shopping in a brand new way—exhibits a great deal of heterogeneity as younger healthier populations search alongside older, sicker customers previously unable to get coverage. While private exchanges can, and do, engage in experimentation, public exchanges, where regulators and political pressures limit the ability to provide different experiences to observationally equivalent consumers, are constrained in their ability to analyze customer behavior and heterogeneity. The ability of private firms to learn

* Cox: Economics Department, UC Berkeley: 530 Evans Hall, Berkeley, CA 94720, (email: nmcox@berkeley.edu). Handel: Economics Department, UC Berkeley and NBER, 501 Evans Hall, Berkeley, CA 94730, handel@berkeley.edu. Kolstad: Wharton School, University of Pennsylvania: 3641 Locust Walk, Philadelphia, PA 19104, and NBER (email: jkolstad@wharton.upenn.edu). Mahoney: University of Chicago Booth School of Business: 5807 South Woodlawn Avenue Chicago, IL 60637, and NBER (email: neale.mahoney@gmail.com). We thank eHealth for data access. We thank our discussants Jason Abaluck and Josh Schwartzstein, and Bob Town and Ashley Swanson for helpful discussion in the development of the paper. All errors are our own.

about consumer behavior—either on average or across the population—or target messaging may impact both rates of insurance take-up and welfare.

I. Data and Setting

To test this hypothesis, we gathered detailed micro-data on a large scale experiment conducted by a private sector exchange. The exchange acts as a privately-run alternative to the national health insurance exchange, HealthCare.gov, and many state-based exchanges. All of these different sites allow consumers to run an individualized search (based on age, gender, and location) for insurance policies and view the results in a listed manner. One main difference are the many interfaces and innovative shopping “tools” provided for consumers on the private website. These tools, for example, provide information about important aspects of the ACA, and allow consumers to compare plans, or search for specific doctors within provider networks.

In February and March of 2014, the exchange conducted a randomized experiment in which new users to the site were subjected to either a control experience or one of two different treatment experiences that reminded them about the upcoming enrollment deadline in compliance with the individual mandate under the ACA. The experiment was conducted from 10 days prior to one of the three different deadlines to 4 days prior, at which point the experience with the greatest success (as measured by conversion rate) was implemented completely. In particular, the first campaign was run from February 5-11 (leading up to the February 15th enrollment deadline), the second from March 5-10 (leading up to the March 15th deadline), and the third from March 21-26 (leading up to the March 31st deadline).

Figure 1 displays examples of the control and a typical treatment homepage. The control experience was the normal homepage that contained no information about the enrollment deadline or related penalties. The two treatment experiences both featured a real-time “countdown clock” and

emphasized the relevant deadline. Treatment B was run in all three campaigns. It consists of a countdown clock, the message “Time is running out” and a sub-message that contained the date of the enrollment deadline. In one campaign it also included the wording “Avoid tax penalties.” Treatment C was run only in the first and second campaigns. Like Treatment B, it contained a countdown clock. However it featured the message “Get Affordable Health Insurance” and a sub-message such as “Apply before March 15th for April 1st coverage.” Because the two treatments are similar in that they both have a countdown clock, we conduct analyses where we pool them as a single treatment and where we examine both of the treatment effects separately.

For each day of each campaign, we observe how many individuals of each demographic group (gender \times age \times family size) and state of residence visited the site and were channeled to one of the three homepages. Therefore, our data is not individual specific, but rather aggregated to day \times demographic group \times state \times treatment level “cells.” We also observe at this level how many visitors submitted applications for health insurance on the site, allowing us to calculate a cell-specific “conversion rate”, defined as the number of submitted applications divided by the total number of unique visitors.

The Online Appendix shows the demographic summary statistics across campaigns and treatments. Across all three campaigns and treatments, we observe a total of 662,713 unique visitors to the site; however, only 310,210 of these observations contain demographic information, and these individuals were more likely to have submitted an application. The table confirms the randomized nature of the experiment. Regardless of treatment, visitors are on average near 40 years old, slightly more likely to be female, and have 1.6 family members. The Online Appendix displays the distribution of users in the experiment by self-reported age and gender. Most users are in the range of 20-60 years old, and both genders seem well represented across age groups. We also look at whether the

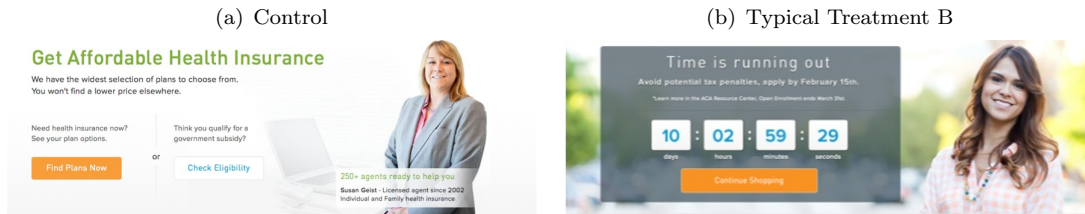


FIGURE 1. HOMEPAGE EXPERIENCES

composition and behavior of visitors to the site changes as we approach deadlines. The Online Appendix shows (using data pooled across treatments and campaigns) that the average visitor tends to be older and has a higher likelihood of submitting an application as we approach the enrollment deadline.

II. Empirical Approach

In this section we outline our approach to estimating impacts of the different messaging treatments and the increase in conversions from the optimal targeting scheme. Below, we focus on the case with a single pooled treatment. In the Online Appendix we discuss how to extend our approach to the multiple treatment case we also analyze.

Let i denote the day \times demographic group \times state \times treatment cell and g denote the “group” at which we allow for heterogeneous treatment effects. For instance, if we want to allow for heterogeneity by gender, we could specify two groups: males and females. To estimate the impact of messaging for a given group g , we estimate the following model:

$$r_{ig} = \alpha_g + \beta_g T_{ig} + X'_{ig} \delta_g + \epsilon_{ig}$$

where r_{ig} is the conversion rate, α_g is a group-specific fixed effect, T_{ig} is the pooled treatment indicator that takes on a value of 1 if the cell receives either of treatments and value of 0 otherwise, and X_{ig} is a set of controls for demographic characteristics that vary at the cell level. The impact of the treatment is given by β_g , which captures the differential impact of the pooled treatment on conversion rates relative to the average conversion rate for the control group. Be-

cause our treatment is randomly assigned, the covariates X_{ig} are not required to recover an unbiased estimate of β_g but help increase the precision of the estimate.

Using the estimated treatment effects β_g , we are interested in estimating the impact of an “optimal” targeting scheme.¹ If we knew the exact values of the β_g ’s, we could recover the impact of targeting by assigning the treatment to a given group if and only if it increased conversions relative to the control (i.e., $\beta_g > 0$). The counterfactual treatment effect from optimal targeting r^* would be given by

$$r^* = \sum_g s_g \max\{\beta_g, 0\}$$

where s_g is the fraction of the population in demographic group g .

However, in any non-infinite sample, we cannot recover the true treatment effect, but instead we recover an *estimate* of its value $\hat{\beta}_g$. If we plug our estimates into the equation above, we will overestimate the gains from targeting because we will sometimes estimate a positive treatment effect even if the true value is negative. We therefore need to adjust our estimates to account for this source of upward bias.

We construct an adjusted estimate of the impact of targeting with

$$\hat{r}^* = \sum_g s_g 1(\hat{\beta}_g > 0) \left\{ \hat{\beta}_g - E[\epsilon_{\beta_g} \mid \hat{\beta}_g > 0] \right\}$$

¹We note that this is only optimal in the sense that we use the observables captured but does not reflect the universe of approaches. While very interesting, computing how to best undertake both learning and targeting more generally or in the case of an insurance exchange is beyond the scope of this paper.

where $E[\epsilon_{\beta_g} | \hat{\beta}_g > 0]$ is the term that adjusts the estimate. This term can be written as $E[\epsilon_{\beta_g} | \hat{\beta}_g > 0] = \frac{\sigma_{\beta_g}}{\sqrt{n_g}} \lambda\left(\frac{-\hat{\beta}_g}{\sigma_{\beta_g}/\sqrt{n_g}}\right)$, where λ is the inverse Mill's Ratio, defined as $\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)]$, Φ , ϕ are the standard normal CDF and PDF, respectively, and $\sigma_{\beta_g}/\sqrt{n_g}$ is the standard error of $\hat{\beta}_g$.

We construct standard errors for \hat{r}^* using the bootstrap method with 500 iterations. For each iteration, we draw a new set of observations for every group g , estimate a new set of treatment effects $\hat{\beta}_g$, and construct a new measure of \hat{r}^* . In the Online Appendix, we discuss the adjustment formula in more detail, show Monte Carlo simulations, and discuss how to extend our approach to the multiple treatment setting.

III. Value of Targeted Messaging in Insurance Exchanges

Table 1 shows the impact of different optimal targeting schemes. The mean conversion rate in the control group is 4.95 percent. The first section of the table shows the effects of providing the same treatment to all of the groups in the data, and therefore does not exploit any heterogeneity. The average effects of the treatments are modest. The pooled treatment raises the conversion rate by 0.12 percentage points on a base of 4.95 percent and is statistically insignificant. Separating the two, treatment C is slightly larger at 0.22 and statistically distinguishable from zero, while Treatment B is statistically insignificant.

The remainder of the table investigates the impact of targeting the treatments based on heterogeneity in three dimensions: Time targeting allows for different treatments by the number of days before the deadline. Demographic targeting allows for different treatments by gender \times age \times family size group. State targeting allows for different treatments by the consumer's state of residence.

The second section of the table shows the effect of different types of targeting with the single pooled treatment. Time and demographic targeting provides virtually no im-

provement relative to the 0.12 increase in conversions from providing everyone with the pooled treatment. State-specific targeting is more effective, raising the conversion rate by 0.33 percentage points or about 7 percent of the 4.95 baseline. Combining the state, time, and demographic targeting raises the conversion rate by 0.91 percentage points or 18 percent of the control mean. We can rule out an increase below 0.37 percentage points with a 95 percent confidence interval.

The final section of the table shows the effects of different types of targeting, additionally allowing for heterogeneity by Treatment B and Treatment C. Separating the treatments has a fairly minor impact on time and demographic targeting but substantially increases the impact of state specific targeting to 0.61 percentage points. Allowing state, time, and demographic targeting, along with the differences by Treatment B and C has the largest effect, raising conversions by 2.07 percentage points or 41.8 percent of the baseline. In this specification, the smaller samples in each group reduces the precision of this estimate, but we can still rule out an effect below 1.43 percentage points, or 29 percent of the control mean, with a 95 percent confidence interval.

IV. Conclusion

Taken together, our results suggest that the different messages tested by the firm did marginally increase conversions, but optimal targeting across time, demographics, and geography increased conversions substantially. Specifically, by targeting the messages, we estimate conversion rates could be increased more than 29 percent of the control conversion rate. Our findings suggest that limits, either explicit in regulations or implicit due to political pressure, on public exchanges may impact consumer behavior. Assuming consumers are able to choose a plan on an exchange that is preferred to remaining uninsured, our results suggest that not only do these limits impact conversion rates themselves but they may also reduce welfare.

Our study has a number of important

TABLE 1—TOTAL TREATMENT EFFECTS WITH TARGETING

	Treatment Effect (Percentage Points)	95 Percent Confidence Interval
No Targeting:		
<i>Treatment B & C Pooled</i>	0.12	(-0.01 , 0.26)
<i>Treatment B</i>	0.06	(-0.01 , 0.21)
<i>Treatment C</i>	0.22*	(0.05 , 0.39)
Targeting, B & C Pooled:		
<i>Time Targeting</i>	0.14*	(0.02 , 0.33)
<i>Demographic Targeting</i>	0.13*	(0.06 , 0.28)
<i>State Targeting</i>	0.33*†	(0.26 , 0.50)
<i>State, Time, & Demo Targeting</i>	0.91*†	(0.37 , 1.23)
Targeting, B & C Separate:		
<i>Time Targeting</i>	0.15*	(0.09 , 0.36)
<i>Demographic Targeting</i>	0.21*†	(0.19 , 0.43)
<i>State Targeting</i>	0.61*†	(0.52 , 0.88)
<i>State, Time, & Demo Targeting</i>	2.07*†	(1.43 , 2.40)

Note: Control conversion rate is 4.95 percent. Analysis restricted to observations with demographic information (N=305,720).

* denotes that the treatment effect is significantly different from zero at the 95% level.

† denotes that the treatment effect is significantly different from the pooled (B & C), untargeted treatment effect at the 95% level.

limitations. First, we study one specific message in a specific time period: the impact of reminders about deadlines during the first open enrollment for the ACA in which the deadline was being moved. Furthermore, we study a targeting rule based on the data we have available and using ad hoc groupings (e.g above versus below median age). In future work we will model the optimal demographic rules given the data we have available. Beyond that, firms could approach A/B testing not only to measure specific messages in existing populations but to optimize the overall targeting approach. While that is clearly beyond the scope of this paper we believe these results underscore the value of pursuing such a research agenda, both for firms and for researchers and policy makers hoping to improve the experience of consumers purchasing insurance through exchanges.

REFERENCES

- ness: A Large Scale Field Experiment.” *Econometrica*.
- Brown, Brad, Michael Chui, and James Manyika.** 2011. “Are You Ready for the Era of “Big Data”?” *McKinsey Quarterly*, October(4).
- Ericson, Keith, and Amanda Starc.** 2013. “How Product Standardization Affects Choice: Evidence from the Massachusetts Health Insurance Exchange.”
- Goldfarb, Avid.** 2014. “What is Different About Online Advertising?” *Review of Industrial Organization*, 44(4): 115–129.
- Lewis, Randall, and David Reily.** 2014. “Advertising Effectively Influences Older Users: How Field Experiments Can Improve Measurement and Targeting.” *Review of Industrial Organization*, 44(4): 147–154.
- Blake, Timothy, Chris Nosko, and Steve Tadelis.** forthcoming. “Consumer Heterogeneity and Paid Search Effective-