J Econ Sci Assoc (2015) 1:29–42
DOI 10.1007/s40881-015-0005-3

ORIGINAL PAPER

Online field experiments: a selective survey of methods

Yan Chen[1] · **Joseph Konstan**[2]

Received: 1 November 2014 / Revised: 12 March 2015 / Accepted: 16 March 2015 /
Published online: 19 May 2015

Abstract The Internet presents today's researchers with unprecedented opportunities to conduct field experiments. Using examples from Economics and Computer Science, we present an analysis of the design choices, with particular attention to the underlying technologies, in conducting online field experiments and report on lessons learned.

Keyword Online field experiment · A/B testing · Internet

JEL Classification C93 · H41

1 Introduction

Field experiments allow researchers to combine the control of laboratory experiments with some of the ecological validity of field studies. Areas such as medicine (Lohr et al. 1986), economics (Harrison and List 2004), and social psychology (Lerner et al. 2003) have all incorporated field experiments in their research. One of the challenges of field experiments, however, is the substantial cost of conducting them, particularly at a sufficient scale for studying high-variance social phenomena. Online communities present a more practical and cost effective venue for conducting field experiments. Given sufficient access to a community of

✉ Yan Chen
yanchen@umich.edu

[1] University of Michigan, Ann Arbor, MI, USA

[2] Department of Computer Science and Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455, USA

users and the software substrate for their community, researchers can study both short- and long-term effects of a wide range of manipulations.

In this paper, we present an analysis of the design choices for online field experiments using representative studies from both Economics and Computer Science. Within Computer Science, we focus on two subfields, i.e., Human–Computer Interactions (HCI), and Computer-Supported Collaborative Work (CSCW). We first summarize current methods for conducting online field experiments, with particular emphasis on the underlying technologies, and then offer some insights and design advice for social scientists interested in conducting such studies.

From the extensive catalog of online field experiments, we choose a representative set of academic studies which use a variety of technologies and cover a broad spectrum of online sites, including those focused on social networking (Facebook, LinkedIn), user-generated content (Wikipedia, MovieLens), e-commerce (eBay, Yahoo!), online games (World of Warcraft), crowdfunding (Kiva), and crowdsourcing (Google Answers, TopCoder, oDesk, Taskcn). Note that we do not include experiments conducted on Amazon's Mechanical Turk, as they have been covered in a separate survey (Horton et al. 2011). Nor do we include experiments conducted on platforms designed for behavioral experimentation, such as LabintheWild (Reinecke and Gajos 2015) and TestMyBrain (Germine et al. 2012).

We also note that IT companies, including Amazon, eBay, Facebook, Google, LinkedIn, Microsoft, Netflix, ShopDirect, Yahoo, and Zynga, conduct a large number of commercial online field experiments, sometimes called A/B testing, to evaluate new ideas, guide product development, and improve interface design.[1] Although the vast majority of these experiments are not intended for publication and are thus not discussed here, the veracity of these studies nonetheless primarily depends on the same methodological issues academic researchers are concerned with that we discuss in this paper.

2 Technologies for intervention

In this section, we discuss four basic experimental technologies for intervention within an online community: email and SMS/texting, modified web interfaces, bots, and add-ons. For each technology, we provide case studies to demonstrate how the underlying technology has been used for intervention.

2.1 Email and text

Email is one of the most common intervention technologies used by researchers. Compared to modified web interface, email is more likely to get participant

[1] Google alone conducts more than 10,000 online field experiments per year (private communication with Hal Varian). At Microsoft's Bing, over 200 concurrent experiments are running on any given day, involving about 100 million active monthly customers (Kohavi et al. 2013).

attention. To illustrate the use of email as a tool for intervention, we examine studies in user-generated content and crowdfunding.

In the first study, Ling et al. (2005) conduct four field experiments with members of the MovieLens online movie recommender community (http://www.movielens.org). In three of these experiments, selected users of the system receive an email message asking them to rate more movies (i.e., to contribute effort and knowledge to the community). In all, over 2400 users receive an email message crafted to test hypotheses based on the Collective Effort Model from social psychology (Karau and Williams 1993). These experiments yield several interesting results. First, the researchers find that highlighting a member's uniqueness by pointing out that the member had rated movies rarely rated by others increases rating behavior. Second, they find that setting specific rating goals (either for individuals or for a group) also increases rating behavior. Surprisingly, highlighting the benefits of movie ratings, either to the member or to others, does not increase the number of ratings.

This experiment demonstrates how to conduct an email intervention. It also underscores the importance of proper controls in an online field experiment. Rating activity peaked after the mailings, but also after the post-experiment thank-you email. Indeed, they find that any reminder about the site seems to promote more visits. In general, this shows it is good practice to have two control conditions in online field experiments, one without any contact from the experimenters and a placebo condition in which participants are contacted but do not receive any treatment content.

Our second example is from a recent field experiment on Kiva (http://www.kiva.org), the first microlending website to match lenders with entrepreneurs in developing countries. In this study, Chen et al. (2015) run a large-scale randomized field experiment ($n = 22{,}233$) by posting team forum messages.[2] Kiva lenders make zero-interest loans to entrepreneurs in developing countries, often out of pro-social motives (Liu et al. 2012). A unique mechanism to increase lender engagement is the Kiva lending team, a system through which lenders can create teams or join existing teams. A team leaderboard sorts teams by the total loan amounts designated by their team members. To understand the effects of the lending teams mechanism on pro-social lending, the researchers examine the role of coordination in reducing search costs and the role of competition through team goal setting. Compared to the control, they find that goal-setting significantly increases lending activities of previously inactive teams. In their experimental design, Chen et al. use a built-in feature in Kiva to summarize daily forum messages into one email that is sent to each team member's inbox. Thus, their experimental intervention is incorporated into the normal flow of emails that lenders receive.

To prepare for an online field experiment, it is often useful to analyze site archival data through a public application programming interface (API), which enables researchers to download data the site collects about its users. For example,

[2] Founded in 2005, Kiva partners with microfinance institutions and matches individual lenders from developed countries with low-income entrepreneurs in developing countries as well as in selected cities in the United States. Through Kiva's platform, anyone can make a zero-interest loan of $25 or more to support an entrepreneur. As of January 2015, more than 2 million lenders across 208 countries have contributed $666 million in loans, reaching over 1.5 million borrowers in more than 73 countries.

through their empirical analysis of the Kiva archival data, Chen et al. are able to assess the role of teams in influencing lending activities, information which provides guidance for the design of their subsequent field experiment.

Similar to email interventions, text messages have been used effectively to implement field experiments. In developing countries in particular, since cell phone penetration has far exceeded that of the personal computer, texting may be a better method of interventions. Compared to emails, the unique challenge of text messaging is the character limit, as a text message should be concise enough to fit a cell phone screen. We refer the reader to Kast et al. (2011) as an example of a field experiment using text messages to encourage savings among low-income micro-entrepreneurs in Chile.

2.2 Modified web interface

Another technology utilized in online field experiments is the modified web interface. In particular, randomized experiments through modified web interface are often used in the technology industry to evaluate the effects of changes in user interface design. Software packages, such as PlanOut,[3] have been developed to facilitate such experimentation (Bakshy et al. 2014). We examine how modified web interfaces have been used in settings such as ad placement and online employment.

In a large-scale field experiment on Yahoo!, Reiley et al. (2010) investigate whether the competing sponsored advertisements placed at the top of a webpage (*north ads*), exert externalities on each other. To study this question, they run a field experiment on Yahoo, where they randomize the number of north ads from zero to four for a representative sample of search queries. Two experiments were conducted with about 100,000 observations per treatment among Yahoo! Search users. Interestingly, the researchers find that rival north ads impose a *positive* externality on existing north listings. That is, a topmost north ad receives more clicks when additional north ads appear below it. This experiment uses modified web interface to determine user behavior in a domain where existing theory has little to say, but companies care a great deal about.

In a social-advertising experiment, Bakshy et al. (2012) use a modified web interface to investigate the effect of social cues on consumer responses to ads on Facebook. In their first experiment ($n = 23{,}350{,}087$), the researchers provide one to three social cues in word-of-mouth advertising, and then measure how responses increase as a function of the number of cues. In their second experiment ($n = 5{,}735{,}040$), they examine the effect of augmenting ad units with a minimal social cue about a single peer. Their findings show that a social cue significantly increases consumer clicks and connections with the advertised entity. Using a measurement of tie strength based on the total amount of communication between subjects and their peers, they find that these influence effects are greatest for strong ties. Their field experiment allows them to measure the magnitude of effects predicted by network theory.

[3] PlanOut is an open source software package developed by Facebook researchers. For detailed information, see https://facebook.github.io/planout/.

More recently, Gee (2014) presents a field experiment which varies the amount of information seen by 2 million job seekers when viewing 100,000 job postings on LinkedIn (https://www.linkedin.com/job/). Users are randomized into a treatment group who see the true number of people who previously started an application, and a control group who see no such information for any of the postings during the 16 days of the experiment. The findings show that the additional information in the treatment increases the likelihood a person will start and finish an application by 2 to 5 percent. Furthermore, Gee finds that the treatment increases the number of female applicants, a finding of interest to the advertising firms in the high tech and finance industry, where women are under-represented. In this case, the researcher brings her theoretical knowledge and academic perspective to an existing randomized field experiment designed and conducted by a company to gain richer insights.

As a tool for intervention, modified web interface can also be used in combination with emails. For example, Chen et al. (2010a) design a field experiment on MovieLens that sends 398 users a personalized email newsletter, with either social or personal information. Each newsletter contains the same five links: (1) rate popular movies, (2) rate rare movies, (3) invite a buddy to use MovieLens, (4) help us update the MovieLens database, and (5) visit the MovieLens home page. Participants who visit MovieLens during the month after receiving the newsletter receive a slightly modified interface with the four links from the email newsletter included in the "shortcuts" pane of the main MovieLens interface. The authors find that users receiving behavioral information about a median user's total number of movie ratings demonstrate a 530 % increase in their number of ratings if they are below the median. They also find that users who receive the average user's net benefit score increase activities that help others if they are above average.

From a design standpoint, this study follows user behavior for an extended period of time, which enables the experimenters to detect whether the effects are long lasting or temporal substitution. To correctly estimate the effects of an intervention, the experimenter should consider temporal substitution or spatial displacement, whichever is appropriate.[4] This study contributes to our understanding of the effects of social information interventions.

2.3 Bots

Another technology available for online field experiments is the bot, a program or script that makes automated edits or suggestions. Wikipedia is one community that allows bots if the experimenter receives approval from a group of designated Wikipedia users with the technical skills and wiki-experience to oversee and make decisions on bot activity.[5] Bots on Wikipedia are governed by the following policy,

[4] An example of spatial displacement in the blood donation context is reported in Lacetera et al. (2012). The authors find that, while economic incentives increase blood donations, a substantial proportion of this increase is explained by donors leaving neighboring drives without incentives to attend those with incentives.

[5] Wikipedia's bot policy can be found at https://en.wikipedia.org/wiki/Wikipedia:Bot_policy. The approval procedure can be found at https://en.wikipedia.org/wiki/Wikipedia:Bots/Requests_for_approval.

"The burden of proof is on the bot-maker to demonstrate that the bot is harmless, is useful, is not a server hog, and has been approved" by the Bot Approvals Group.

One study that makes use of a bot is that of Cosley et al. (2007), who deploy an intelligent task-routing agent, SuggestBot, to study how Wikipedia workload distribution interfaces affect the amount of work members undertake and complete. They deploy SuggestBot to pre-process a dump of Wikipedia data to build a learning model of what articles a user might be interested in editing based on their past editing behavior. SuggestBot then recommends editing jobs to users through their talk pages. Their findings show that personalized recommendations lead to nearly four times as many actual edits as random suggestions.

One challenge in using bots on a third-party website is the level of detail of observation available (e.g., observation of edits, but not reading behavior), but this is all determined by the nature of the programming interface for extending the underlying site or browser to implement monitoring.[6] Nonetheless, bots can be used to address technical design questions motivated by social science. They can also be a way to enhance matching at a relatively low cost.

2.4 Add-ons

A final technology that can be utilized by researchers is the add-on, such as a browser extension,[7] that can monitor a participant's behavior across multiple sites. For example, once users install the MTogether browser extension or mobile app, researchers can access their cross-site behavior over an extended period of time, creating a large-scale longitudinal panel to facilitate data collection and intervention (Resnick et al. forthcoming).

The advantage of a browser extension is significant when the experimental intervention is based on information gathered across multiple sites. For example, Munson et al. (2013) deploy a browser extension to nudge users to read balanced political viewpoints. The extension monitors users' web browsing, accesses and classifies their browsing history, dynamically aggregates the political leaning of their reading selections, and then encourages those whose reading leans one way or the other to read a more balanced selection of news. Users see a balance icon that indicates their leaning as well as recommendations of sites representing more neutral positions or the "other" side. Users in the control group receive aggregate statistics only after the 28th day of the experiment. Compared to the control group, users in the treatment show a modest move toward balanced exposure. This study provides a practical tool to potentially alleviate the polarization of the US political landscape.

Another advantage of an add-on is that interventions can be carried out in real time. For example, in response to earlier research showing that being reverted (and

[6] For example, Chrome, Firefox and Internet Explorer all have interfaces where researchers can extend the browser to monitor page views and scrolling activities. In some cases, researchers can extend underlying sites through a programming interface that may permit notice of reading or editing behavior.

[7] A browser extension is a computer program that extends the functionality of a web browser, such as improving its user interface, without directly affecting viewable content of a web page. Source: https://msdn.microsoft.com/en-us/library/aa753587(VS.85).aspx.

in many cases being reverted without comment or rudely) is a cause for attrition among new Wikipedia editors, Halfaker et al. (2011) deploy an add-on (built in JavaScript) to alert Wikipedia editors who are performing revert operations that they are reverting a new editor. It also provides a convenient interface for sending an explanatory message to that new editor. This intervention has led to significant changes in interaction and an increase in retention of new editors (based on the messaging and warning, respectively). It provides a much needed technology to increase the retention of new Wikipedia editors, which can be extended to other online communities as well.

In sum, the technology available within online communities provides researchers with the opportunity to conduct large-scale and real-time interventions that better capture participant behavior in field experiments.

3 Design choices

To aid researchers interested in conducting an online field experiment, we outline a number of design considerations, including (1) the access and degree of control the experimenter exerts over the online venue, (2) issues of recruiting, informed consent, and the IRB, (3) the identifiability and authentication of subjects, and (4) the nature of the control group. Note that these dimensions exclude the three core design features of any experiment—the hypotheses, the experimental conditions, and the presentation of the experimental manipulation, as these vary substantially with each individual study. We also do not include power analysis as this can be found in statistics textbooks.

3.1 Access and degree of control

When a researcher has the flexibility to choose among several different sites on which to conduct a study, the degree of experimenter control is an important factor to consider.

1. *Experimenter-as-user* involves minimal or no collaboration with the site owners. On many online sites, experimenters may establish identities as users for the purposes of both gathering field data and introducing interventions. Naturally, both the types of manipulation possible and the data that can be gathered are limited by the system design. Furthermore, some online communities have usage agreements or codes of conduct that prohibit such research uses. The experimenter-as-user approach has been used since the first economic field experiment conducted over the Internet, where Lucking-Reiley (1999) auctioned off pairs of identical collectible Magic: the Gathering trading cards using different auction formats to test the revenue equivalence theorem. Using an Internet newsgroup exclusively devoted to the trading of cards, with substantial trading volume and a variety of auction mechanisms, he found that the Dutch auction produced 30-percent higher revenues than the first-price auction. These results are counter to well-known theoretical predictions and previous laboratory results,

which might be due to several differences between the field setting and those of the laboratory: (1) simultaneous auction for multiple items, (2) real versus induced values, and (3) slow Dutch auctions. A subsequent lab experiment shows that (3) matters—the slower the speed of the Dutch auction, the more revenue it raises (Katok and Kwasnica 2008). This pair of studies demonstrate the complementarity between laboratory and field experiments.

In another study, Resnick et al. (2006) conducted a field experiment on eBay to study Internet reputation systems. In their design, a high-reputation, established eBay seller sold matched pairs of vintage postcards under his regular identity as well as under seven new seller identities (also operated by him). With this approach, they were able to measure the difference in buyers' willingness-to-pay, and put a price on good reputation. Since eBay was not involved in the experiment, data were collected directly from the eBay webpage using a Web crawler, an Internet bot that systematically browses and copies webpages for indexing and data collection. The significance of this experiment is their empirical estimate for a widely discussed parameter, the value of reputation.

Similarly, the experimenter-as-employer model has been used for crowdsourcing experiments, testing social preferences on the now-defunct Google Answers (Harper et al. 2008; Chen et al. 2010b), labor market sorting on TopCoder (Boudreau and Lakhani 2011), and contest behavior on Taskcn (Liu et al. 2014). In one such experimenter-as-employer study, Pallais (2014) evaluates the effects of employment and feedback on subsequent employment outcomes on oDesk (https://www.odesk.com/), an online labor market for freelance workers. In this study, 952 randomly-selected workers are hired for data entry jobs. After job completion, each receives either a detailed or coarse public evaluation. Using oDesk administrative data, Pallais finds that both the act of hiring a worker and the provision of a detailed evaluation substantially improve a participant's subsequent employment rates, earnings and reservation wages. Her results have important public policy implications for markets for inexperienced workers as well as reputation building.

2. *A site with a public interface* is another option that allows for substantial experimenter control. Facebook, LinkedIn, and Wikipedia all encourage the integration of third-party applications. For example, Cosley et al. (2007) use the Wikipedia data dumps to build a model of users (based on editing behavior) and articles to identify the articles a user might be interested in editing. They then deploy SuggestBot to recommend articles to potential editors. Their study illustrates the challenges of working through an open interface, as their profiles are limited to those with existing editing experience. In the online gaming area, Williams et al. (2006) use a public interface to study the social aspect of guilds in World of Warcraft. They gather their data through player bots, interfaces that provide statistics on currently active players.
3. *A collaborative relationship with a site owner* is another choice that can provide a fair amount of data and control. For example, Chen et al. (2006) worked with the Internet Public Library (IPL) to test the effectiveness of various fund-raising mechanisms proposed in the literature. These were implemented through a variety of solicitation delivery interfaces (e.g., pop-up messages, pop-under

messages, and in-window links). Their findings show that Seed Money (i.e., an initial large donation) and Matching mechanisms (i.e., a benefactor commits to match donations at a fixed rate) each generate significantly higher user click-through response rates than a Premium mechanism, which offers donors an award or prize correlated with the gift size. In this case, their collaboration with the IPL staff allows them to collect micro-behavioral data, such as user click-streams. Such collaborative relationships can be extremely effective, but tend to develop slowly as the site owner gains trust in the collaborating researcher. As such they are best viewed as a substantial investment in research infrastructure rather than as a quick target for a single study. Finally, a variation of the collaborative model is to partner with companies through shared research projects that involve doctoral students as interns. Furthermore, many IT companies have been hiring academics, who conduct online field experiments both for the company and for pure academic research.

4. Lastly, *owning your own site* is the option that gives the experimenter the most control and flexibility in the experimental design and data collection. One site, MovieLens, was created by researchers more than a decade ago, and has provided the ability to control and measure every aspect of the system and of user interaction with it over time. For example, it allows researchers to modify the interface, implement varying interfaces for different experimental groups, analyze usage data to assign users into experimental groups, and email users who opt in to invite them to participate in experiments. One study conducted with MovieLens examines the effects of social information on contribution behavior by providing personalized email newsletters with varying social comparison information (Chen et al. 2010a). The experimenters have access to user history data (e.g., number of movies rated, frequency of login, and other usage data) that aids in assigning subjects to groups and in personalizing their newsletters. They were able to track user activity in the month following the newletter mailing (and beyond) to determine the effect of the manipulation on user interaction with the site as a whole. Finally, the site allows for a modified web interface to present the email newsletter links within the site itself. This level of control and observation would be difficult without direct control over the site.

   Despite the advantages, site ownership can be costly. The original MovieLens implementation took about 1 month of development with two masters students working on it. The fixed cost was small because the site was a clone of the EachMovie site that DEC was shutting down, with few features and no design. Since then, the research team has maintained a solid investment in MovieLens, with a full-time staff member supporting its maintenance, ongoing development, and related research—usually working together with two or three part-time undergrads and masters students, and occasionally several Ph.D. students. During an experiment, the costs increase, with a full-time staff member who devotes about 1/4 of his time to site maintenance, a Ph.D. student who devotes about 10 h a week to system development and enhancements, and rotating responsibility in the lab for handling customer support for about one to 2 h per week.

   Starting a site from scratch involves higher fixed costs. For example, launching LabintheWild required 6 months of programming effort. Subsequently, it takes

approximately ten programming hours to maintain the site (excluding the construction of new experiments) and an additional 10 h per week for general maintenance, including writing blog posts, updating its facebook page and answering participant emails.[8]

3.2 Recruiting, informed consent and the IRB

In addition to considering what type of experimenter control is best suited, researchers must consider issues related to subject recruiting and ethical issues related to the experiment. Online field experiments use two types of subject recruiting. The first type is natural selection. In the eBay field experiments discussed above, the experimental tasks are natural tasks that participants interested the item undertake. These are natural field experiments (Harrison and List 2004), where participants do not know that they are in an experiment. In nearly all cases, no informed consent is presented to the participants because awareness of the research study and being monitored can affect behavior (List 2008).

The second type of online recruiting method is sampling. An experimenter with access to a database of site users can generate a pool of potential subjects, and in some way recruit them to participate. From the pool, the experimenter may invite a random sample, may create a stratified or other systematic sample, or may simply implement the experimental interface across the entire pool. In one study, Leider et al. (2009) recruit their subjects from Facebook, but then direct them to the researchers' own website to conduct the actual experiment.

Subject recruitment may be explicit, as in Chen et al. (2010a), who recruit via email, with those who reply as subjects. Other experiments, such as the email studies shown in Ling et al. (2005), randomly select users and assign those users into groups, where being sent the email is the experimental treatment. By contrast, Sen et al.'s (2006) tagging experiments present the interface to the entire community. For experiments which accept convenience samples of those users who volunteer, who visit the site, or who otherwise discover the manipulation, there is the concern of sample selection bias. Even studies that do not require explicit consent, such as Cosley et al. (2007) or Sen et al. (2006), face sample selection biased towards frequent or attentive users.

The recruitment strategy for online field experiments is closely related to the question of informed consent. Compared with laboratory experiments, it is much more common for field experiments to request a waiver of informed consent so as to avoid changing the behavior of the subject.

In general, researcher who plan to run online field experiments should go through the IRB process, to have a disinterested third party evaluate the ethical aspect of the proposed experiment, even though the process might not be able to screen out all unethical studies. In our experience, some university IRBs are reasonable and efficient, while others bureaucratic. In the industry, to our knowledge, Yahoo Research established an IRB process for online field experiments, whereas other

[8] Private communication with Katharina Reinecke.

major IT companies do not, although some have tight privacy controls on all use of individual-level data.

3.3 Identification and authentication

Researchers interested in conducting online field experiments need to consider how they will accurately identify user and track individual activities, as most studies benefit from the ability to identify users over a period of time.

Identification requires that a user offer a unique identifier, such as a registered login name. Authentication is a process that verifies the proffered identity, to increase the confidence that the user proffering the identity is actually the owner of that identity. An identification and authentication system may also ensure that a real-world user has only a single identity in the online community of interest (Friedman and Resnick 2001). Sites that provide personalization or reputation systems typically require login with an ID and password. E-commerce sites may require login, but often not until a purchase is being made. In contrast, many information services, from CNN.com and ESPN.com to the Internet Public Library, do not require any user identification. For these sites, creating an identification system that requires users to create accounts and login might adversely affect usage and public satisfaction with the service, and would therefore likely to be discouraged by the site owners.

Three methods commonly used for tracking users on sites without logins are session tracking, IP addresses, and cookies. Each method has both strengths and weaknesses. For example, session tracking on a web server can identify a sequence of user actions within a session, but not across sessions. IP addresses, on the other hand, can be used to track a user across multiple sessions originating from the same computer. However, they cannot follow a user from computer to computer and are often reissued to a new computer with the original computer receiving a new address. Cookies are small files that a website can ask a user's web browser to store on the user's computer and deliver at a later time. Cookies can identify a user even if her IP address changes, but not if a user moves to a different computer or browser, or chooses to reject cookies.

In one study, Chen et al. (2006) use cookies to ensure that a user remains in the same experimental group throughout the experiment. Users who store cookies receive the same campaign message. For other users, the researchers attempt to write a cookie to create an ID for the user in the experimental database. This approach cannot protect against users returning via multiple machines, but it is a practical solution. We should note that people who reject cookies may be more technologically savvy than the average user, which raises sample bias questions for some studies. In the end, there is no perfect method for determining online identification and authentication. Whenever possible, researchers should conduct online field experiments on sites which require login with an ID and password.

3.4 Control group

Finally, designing appropriate control conditions for online field experiments can be challenging. In many cases, it is necessary to have at least two different control groups. One group receives a carefully matched stimulus, with the exception of the hypothesized active ingredient. For example, if studying personalization, the control group could receive an unpersonalized version of the interface; if studying varying content, the control group could receive the same media, but different content; if studying the medium, the control group could receive the same content, but with a different medium. We call this type of control the *placebo*, as it is similar to the placebo in medical experiments. The placebo design can improve the precision in which the causal effects are estimated (Johnson et al. 2015; chapter 5 in Gerber and Green 2012). However, an online experiment often requires an additional control in which users are not contacted by the experimenters, to help estimate the extent of any Hawthorne effects or mere contact effect (Ling et al. 2005). To be effective, this control needs to be selected from the group of recruits, volunteers, or other eligible subjects.

4 Conclusion

The number of online field experiments has grown tremendously in recent years, spanning fields in economics and computer science as diverse as public finance, labor economics, industrial organization, development economics, human–computer interactions, computer-supported collaborative work, and e-commerce. With the expansion of the Web and e-commerce, we expect this number to grow even more. While some experiments do not lend themselves to Internet testing, we expect that many field experiments on charitable giving, social networks, auctions, personalized advertisement, and health behavior will be conducted online.

Compared to their offline counterparts, online field experiments tend to have both a larger number of observations and natural language as variables of interest, which sometimes require new tools for data manipulation and analysis. We refer the reader to Varian (2014) for an overview of these new tools and machine learning techniques.

Working at the intersection of economics and computer science, this paper has provided a discussion of the main technologies for conducting such experiments, including case studies to highlight the potential for online field experiments. It has also provided insight into some of the design considerations for researchers in navigating the online field experiment arena.

Acknowledgments We would like to thank Eytan Bakshy, Tawanna Dillahunt, Sara Kiesler, Nancy Kotzian, Robert Kraut, David Reiley, Katharina Reinecke, Paul Resnick, John Riedl and Loren Terveen, for helpful conversations on the topic and comments on a previous version. We are grateful to Robert Slonim and two anonymous referees for their comments and suggestions which significantly improve the paper. The financial support from the National Science Foundation through Grants Nos. IIS-0325837 and BCS-1111019 is gratefully acknowledged. Chen: School of Information, University of Michigan, 105 State Street, Ann Arbor, MI 48109-2112. Email: yanchen@umich.edu. Konstan: Department of Computer

Science and Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455. Email: konstan@cs.umn.edu.

References

Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). Designing and deploying online field experiments. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 ACM New York, NY, USA, pp. 283–292.

Bakshy, E., Eckles, D., Yan, R., & Rosenn, I. (2012). Social influence in social advertising: Evidence from field experiments. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12 ACM New York, NY, USA, pp. 146–161.

Boudreau, K. J., & Lakhani, K. (2011). The confederacy of heterogeneous software organizations and heterogeneous developers: field experimental evidence on sorting and worker effort. doi:10.2139/ssrn.1898277

Chen, Y., Li, X., & MacKie-Mason, J. (2006). Online fund-raising mechanisms: A field experiment. *Contributions to Economic Analysis and Policy*, Berkeley Electronic Press, *5*(2), Article 4.

Chen, Y. F. M. H., Konstan, J., & Li, S. X. (2010a). Social comparisons and contributions to online communities: A field experiment on movielens. *American Economic Review*, *100*(4), 1358–1398.

Chen, Y., Teck-Hua, H., & Kim, Y.-M. (2010b). Knowledge market design: A field experiment at Google Answers. *Journal of Public Economic Theory*, *12*(4), 641–664.

Chen, R., Chen, Y., Liu, Y., & Mei, Q. (2015). Does team competition increase pro-social Lending? Evidence from online microfinance. *Games and Economic Behavior*. doi:10.1016/j.geb.2015.02.001.

Cosley, D., Frankowski, D., Terveen, L., & Riedl, J. (2007). SuggestBot: Using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pp. 32–41. Downloaded on February 23, 2003 at http://www.communitytechnology.org/nsf_ci_report/.

Friedman, E. J., & Resnick, P. (2001). The social cost of cheap pseudonyms. *Journal of Economics and Management Strategy*, *10*(2), 173–199.

Gee, L. K. (2014). The More You Know: Information Effects in Job Application Rates by Gender in A Large Field Experiment. Tufts University Manuscript.

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. New York: WW Norton & Company, Inc.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin and Review*, *19*(5), 847–857.

Halfaker, A., Song, B., Stuart, D. A., Kittur, A., & Riedl, J. (2011). NICE: Social translucence through UI intervention. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11 ACM New York, NY, USA, pp. 101–104.

Harper, F. M., Raban, D., Rafaeli, S., & Konstan, J. A. (2008). Predictors of answer quality in online Q&A sites. In *CHI '08: Proceeding of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*, ACM New York, NY, pp. 865–874.

Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*(4), 1009–1055.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425.

Johnson, G. A., Lewis, R. A., & Reiley, D. (2015). Location, location, location: repetition and proximity increase advertising effectiveness. http://www.davidreiley.com/papers/LocationLocationLocation.pdf.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*, 681–706.

Kast, F., Meier, S., & Pomeranz, D. (2011). Under-savers anonymous: Evidence on self-help groups and peer pressure as a savings commitment device. Working Paper, Columbia Business School.

Katok, E., & Kwasnica, A. M. (2008). Time is money: The effect of clock speed on seller's revenue in Dutch auctions. *Experimental Economics*, *11*(4), 344–357.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '13 ACM New York, NY, USA, pp. 1168–1176.

Lacetera, N., Macis, M., & Slonim, R. (2012). Will there be blood? Incentives and displacement effects in pro-social behavior. *American Economic Journal: Economic Policy*, *4*(1), 186–223.

Leider, S., Mobius, M. M., Rosenblat, T., & Do, Q.-A. (2009). Directed altruism and enforced reciprocity in social networks: How much is a friend worth? *Quarterly Journal of Economics*, *124*(4), 1815–1851.

Lerner, J. S., Gonzalez, R. M., Small, D. A., & Fischhoff, B. (2003). Effects of fear and anger on perceived risks of terrorism a national field experiment. *Psychological Science*, *14*(2), 144–150.

Ling, K., Beenen, G., Ludford, P., Wang, X., Chang, K., Li, X., et al. (2005). Using social psychology to motivate contributions to online communities. *Journal of Computer-Mediated Communication*, *10*(4). doi:10.1111/j.1083-6101.2005.tb00273.x.

List, J. A. (2008). Informed consent in social science. *Science*, *322*, 672.

Liu, T. X., Yang, J., Adamic, L. A., & Chen, Y. (2014). Crowdsourcing with all-pay auctions: a field experiment on Taskcn. *Management Science 60*(8), 2020–2037.

Liu, Y., Chen, R., Chen, Y., Mei, Q., & Salib, S. (2012). "I loan because...": Understanding motivations for pro-social lending. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12 ACM New York, NY, USA, pp. 503–512.

Lohr, K. N., Brook, R. H., Kamberg, C. J., Goldberg, G. A., Leibowitz, A., Keesey, J., et al. (1986). Use of medical care in the RAND health insurance experiment: Diagnosis-and service-specific analyses in a randomized controlled trial. *Medical Care 24*(9 Suppl), S1–S87.

Lucking-Reiley, D. (1999). Using field experiments to test equivalence between auction formats: Magic on the internet. *American Economic Review*, *89*(5), 1063–1080.

Munson, S., Lee, S., & Resnick, P. (2013). Encouraging reading of diverse political viewpoints with a browser widget. In *International AAAI Conference on Weblogs and Social Media*, ICWSM 2013, Boston, USA.

Pallais, A. (2014). Inefficient hiring in entry-level labor markets. *American Economic Review*, *104*(11), 3565–3599.

Reiley, D. H., Li, S.-M., & Lewis, R. A. (2010). Northern exposure: A field experiment measuring externalities between search advertisements. In *Proceedings of the 11th ACM Conference on Electronic Commerce*, EC'10 ACM New York, NY, USA, pp. 297–304.

Reinecke, K., & Gajos, K. (2015). LabintheWild: Conducting large-scale online experiments with uncompensated samples. In *Computer supported cooperative work and social computing (CSCW)*, Vancouver, BC, Canada.

Resnick, P., Adar, E., & Lampe, C. What social media data we are missing and how to get it. *The Annals of the American Academy of Political and Social Science* (forthcoming).

Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2006). The value of reputation on eBay: A controlled experiment. *Experimental Economics*, *9*(2), 79–101.

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., et al. (2006). Tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, ACM, pp. 181–190.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, *28*(2), 3–27.

Williams, D., Ducheneaut, N., Xiong, L., Zhang, Y., Yee, N., & Nickell, E. (2006). From tree house to barracks the social life of guilds in world of warcraft. *Games and Culture*, *1*(4), 338–361.