Facebook as a Research Tool for the Social Sciences

Opportunities, Challenges, Ethical Considerations, and Practical Guidelines

Michal Kosinski
Stanford University

Sandra C. Matz
University of Cambridge

Samuel D. Gosling
University of Texas and University of Melbourne

Vesselin Popov and David Stillwell
University of Cambridge

Facebook is rapidly gaining recognition as a powerful research tool for the social sciences. It constitutes a large and diverse pool of participants, who can be selectively recruited for both online and offline studies. Additionally, it facilitates data collection by storing detailed records of its users’ demographic profiles, social interactions, and behaviors. With participants’ consent, these data can be recorded retrospectively in a convenient, accurate, and inexpensive way. Based on our experience in designing, implementing, and maintaining multiple Facebook-based psychological studies that attracted over 10 million participants, we demonstrate how to recruit participants using Facebook, incentivize them effectively, and maximize their engagement. We also outline the most important opportunities and challenges associated with using Facebook for research, provide several practical guidelines on how to successfully implement studies on Facebook, and finally, discuss ethical considerations.

Keywords: Facebook, research design, ethics, big data, myPersonality, snowball sampling

Facebook, the largest and most popular online social network, has become a significant part of daily life for nearly 1.4 billion people around the world (Facebook Inc., 2015). An increasing number of studies focus on Facebook’s influence on individuals and societies (for a comprehensive review, see Wilson, Gosling, & Graham, 2012). However, its potential as a powerful research tool—for both online and offline research in the social sciences—has been largely overlooked. This is unfortunate, because Facebook provides a number of tools that can be used to inexpensively recruit large and diverse samples. These tools could help to address a major challenge in social science: its overreliance on relatively small, student,1 and disproportionately Western, educated, industrialized, rich, and democratic samples (Henrich, Heine, & Norenzayan, 2010). Furthermore, Facebook can be used as a powerful data-recording tool because it stores detailed demographic profiles and records of an enormous amount of actual behavior expressed in a natural environment. Investigators can, with users’ consent, record their data retrospectively, greatly reducing the limitations associated with self-reported and laboratory-based studies (Paulhus & Vazire, 2007).

In this article, we review the distinct opportunities and challenges offered by Facebook to researchers, provide a number of practical recommendations for effectively conducting research within this environment, and discuss several ethical considerations. We argue that using Facebook in research is often relatively straightforward and generally produces robust results. Contrary to a widespread belief, neither the development of a dedicated Facebook app nor substantial changes to existing research procedures are prerequisites for researchers to benefit from Facebook features in their study designs. Doing so could be as easy as posting an advert on Facebook targeting a specific subpopulation or replacing an online or offline demographic survey with a button that enables participants to donate their behavioral and demographic data (stored on a Facebook profile) to the researchers.

However, to fully tap into the large samples and rich data offered by Facebook, researchers may need to rethink traditional research designs and acquire new skills, including more advanced statistical techniques and a sensitivity to front-end Web design. Similarly, because participants can abandon online studies with just one click, greater care must be taken to meet their needs and optimize their experience in the time available. Finally, some of Facebook’s advantages (such as easy access to large amounts of personal data) introduce serious ethical challenges that must be taken to meet their needs and optimize their experience in the time available. Finally, some of Facebook’s advantages (such as easy access to large amounts of personal data) introduce serious ethical challenges that must be taken to meet their needs and optimize their experience in the time available.

References

Michal Kosinski, Graduate School of Business, Stanford University; Sandra C. Matz, The Psychometrics Centre, University of Cambridge; Samuel D. Gosling, Department of Psychology, University of Texas, and Melbourne School of Psychological Sciences, University of Melbourne; Vesselin Popov and David Stillwell, The Psychometrics Centre, University of Cambridge.

myPersonality application is owned by David Stillwell and Michal Kosinski. We thank Lindsay Graham, Robert Wilson, Kristin Laurin, and Wu Youyou for their feedback on this article.

Correspondence concerning this article should be addressed to Michal Kosinski, Stanford Graduate School of Business, Knight Management Center, Stanford University, 655 Knight Way, Stanford, CA 94305. E-mail: michalk@stanford.edu

1 Some estimates indicate that 85% of psychological studies are based on exclusively undergraduate samples (Gosling et al., 2004).

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
have yet to be addressed in a pragmatic manner by the applicable legal and ethical guidelines.

This article is not a complete compendium of Facebook research methods, but we hope that it will encourage others to consider using Facebook as a research tool, benefiting from our experience and avoiding our mistakes. We draw heavily upon our own experience in designing and maintaining several Facebook studies that have attracted more than 10 million participants. The most popular of our projects, myPersonality, was set up by David Stillwell in 2007 (Stillwell & Kosinski, 2015). It offered participants access to 25 psychological tests, ranging from the 300-item International Personality Item Pool (IPIP) proxy for the NEO-PI-R inventory, to the Satisfaction With Life Scale and a computer-adaptive proxy of Raven’s Standard Progressive Matrices. The participants received immediate feedback on their results and could volunteer to share their Facebook profiles with us; approximately 30% of them (i.e., over 2 million) decided to do so. Recognizing the generosity of the participants who donated their data for research purposes, with their consent, we have shared this resource with the academic community. The Web site, http://mypersonality.org, contains dozens of cleaned, pre-processed, and anonymized databases, a knowledge base, and examples of syntax useful in analysis. Nearly 200 researchers from over 100 academic institutions have used myPersonality data in their work.

The mere fact that Facebook has proven to be the most popular social network since April 2008 does not guarantee that it will continue to hold this title. Friendster, its de facto predecessor, was overtaken by Myspace, which was itself surpassed by Facebook (CNET, 2014). Nevertheless, although the platforms may replace one another, we predict that online social networking as a phenomenon will continue far into the future. It therefore follows that, while some of our Facebook-specific recommendations might become outdated, many others will remain applicable to whichever platform takes its place.

**Recruiting and Retaining Participants**

The size and reach of the Facebook platform offers researchers an unprecedented opportunity to acquire large and diverse samples of participants. Evidently, the Facebook population is not perfectly representative; its users tend to be younger, better educated, and some groups might be entirely excluded (e.g., the Amish, people without Internet access, or those residing in countries that block Facebook on political or religious grounds). However, the sheer size of Facebook’s population implies that even the underrepresented populations are relatively large. For example, as of 2014, nearly 35% of America’s older adults (over 65 years of age) were on Facebook, and their number was quickly growing (Duggan, Ellison, Lampe, Lenhart, & Madden, 2015; Pew Research Center, 2014). Furthermore, Facebook’s global reach supports the investigation of cultural differences that are often overlooked in traditional studies (Heine, Lehman, Peng, Greenholtz, & Of, 2002). For instance, Facebook samples have been used to examine cultural differences in self-presentation in photographs (Huang & Park, 2013), as well as the establishment and maintenance of friendship networks (Peters, Winschiers-Theophilus, & Mennecke, 2013).

The following subsections discuss the ways in which Facebook can be used to recruit and retain large and diverse samples of participants.

**Going Viral: Recruiting Participants Using a Snowball Sampling Approach**

One of the least expensive ways of dipping into the Facebook participant pool is by snowball sampling (Goodman, 1961)—convincing Facebook users to invite their friends to join a study. If enough participants do so, the positive feedback loop may lead to self-sustaining studies with a rapid growth in sample size. For instance, the myPersonality project was originally only shared with the author’s 150 Facebook friends, but it went viral and attracted over 6 million participants in 4 years.

However, achieving such virality is far from straightforward. Only the most engaging studies, such as games or interesting psychological questionnaires that provide compelling feedback, are likely to spread effectively through participants’ social networks. This is because there are many other applications and Web sites (often very well-funded) that compete for users’ attention. The popularity of myPersonality was certainly boosted by a novelty factor; it was the first application of its type to appear on Facebook. Similar personality testing applications released in the following years, often more advanced in terms of design and

---

2 Information, such as names and Facebook IDs, were removed from the datasets and replaced by anonymous user IDs.
representative samples, obtained from panel data, are af-
takers recruited on Amazon’s Mechanical Turk (AMT; of undergraduate Psychology students or professional test-
domized samples, instead relying on ad hoc samples, often
2010). Yet most psychological research does not use ran-
recruited into the sample (Kurant, Markopoulou, & Thiran,
more, people with many friends are more likely to be
method can introduce biases (Kurant, Markopoulou, &
Web sites.
and does limit the exposure of posts created by apps and
owners to spend money on paid advertising, Facebook can
Note, however, that in an attempt to encourage application
friends that have similar personality profiles. Many of the
participants share those reports on Facebook and tag their
friends in the description, which attracts new participants.
Note, however, that in an attempt to encourage application
owners to spend money on paid advertising, Facebook can
do and does limit the exposure of posts created by apps and
Web sites.
Important, snowball sampling methods do not meet
the gold standard of randomized sampling, because the
method can introduce biases (Kurant, Markopoulou, &
Thiran, 2011). The first participants (seeds) are likely to
disproportionately affect the composition of the sample,
because people tend to interact with others similar to them-
selves (McPherson, Smith-Lovin, & Cook, 2001). Further-
more, people with many friends are more likely to be
recruited into the sample (Kurant, Markopoulou, & Thiran,
2010). Yet most psychological research does not use ran-
domized samples, instead relying on ad hoc samples, often
of undergraduate Psychology students or professional test-
takers recruited on Amazon’s Mechanical Turk (AMT;
Chandler, Mueller, & Paolacci, 2014). Even supposedly
representative samples, obtained from panel data, are af-
fected by self-selection biases because only certain types of
people sign up to these panels (Blumberg & Luke, 2009).
In this context, Facebook samples provide an inexpen-
sive and relatively high-quality alternative (Baltar & Bru-
et, 2012). For example, the age distribution of the myPer-
sonality snowballed sample is not more biased than
traditional paper-and-pencil studies published in the Jour-
nal of Personality and Social Psychology (Gosling, Vazire,
Srivastava, & John, 2004). Additionally, the size and di-
versity of the Facebook population help to minimize the
disadvantages of snowball sampling. First, given enough
participants, the representativeness of the population can be
improved by weighting. For example, participants over 55
years of age are underrepresented in the myPersonality
sample, but it still contains nearly 100,000 of them (Still-
well & Kosinski, 2015). Second, employing virality based
on an intrinsic motivation means that people share and
participate in a study out of personal interest rather than
financial gain, which is thus likely to yield data of higher
quality (see the section Improving the Quality of the Self-
Reports). Third, social network information provided by
Facebook can be used to control for the fact that the sample
was collected in a non random way. For example, respondent-
driven sampling provides a mathematical model that weights
the sample to compensate for the fact that it was not collected
via a randomized procedure (Salganik & Heckathorn, 2004).
Finally, as discussed in the next section, researchers can
reduce biases by starting with a diverse set of initial partici-
pants (Morgan, 2008).

**Targeting: Recruitment Through Facebook Advertising**

Facebook’s advertising platform provides an alternative to
snowball sampling, albeit at a price. Researchers can create
and publish an advert that promotes their offline study,
Web site, game, online questionnaire, or Facebook applica-
tion. These adverts can be targeted at users that are
defined by a wide range of preferences (e.g., liking “getting
up early in the morning”); behaviors (liking “running”);
and demographic variables including location, education,
language, political views, ethnicity, sexual orientation, in-
come, and many more.

This approach is particularly useful when seeking
participants that are otherwise hard to reach, such as those
stigmatized in the offline world (Mangan & Reips, 2007) or
those who are hesitant to meet researchers face-to-face
(Batterham, 2014). For example, an advert could be tar-
ged at 40- to 50-year-old men of Hispanic origin, who
live in the United States, and who are interested in rela-

---

3 Facebook provides advice and numerous case studies that help to
understand how to make one’s content viral. It also offers robust docu-
dumentation and samples of code ready to be copied and pasted into one’s
4 Such cross-ratings present a number of interesting research oppor-
tunities of their own. See, for example, Youyou, Kosinski, & Stillwell,
2015.
tionships with other men. At the time of writing, the Facebook advertising platform reported 11,000 individuals matching this profile, who could be reached with an estimated cost-per-click of between $0.40 and $0.80.

Moreover, targeted advertising can help protect participants’ anonymity and decrease the number of (potentially sensitive) questions that have to be asked. In the example just presented, only users matching the targeted profile would see and be able to follow the advert in the first place. Consequently, there is no need to identify the participants, record their Facebook data, or pose questions temporarily sensitive) questions that have to be asked. In the example just presented, only users matching the targeted profile would see and be able to follow the advert in the first place. Consequently, there is no need to identify the participants, record their Facebook data, or pose questions related to their sexual orientation, race, gender, or age.

Previous research shows that Facebook advertising can help reduce the costs of targeted participant recruitment for online surveys (Batterham, 2014) and offline studies (Close, Smaldone, Fennoy, Reame, & Grey, 2013), as well as building long-term participant pools (K. J. Johnson, Mueller, Williams, & Gutmann, 2014). Not only did Facebook ads outperform traditional methods, such as postal surveys (Batterham, 2014), but they were also more cost-efficient than Google advertising, online newsletters, and e-mails (Carlini, Safioti, Rue, & Miles, 2015). Based on 10 recent studies (identified on Scopus.com at the time of writing) that reported using the Facebook advertising platform, the average cost-per-participant was $13.75, with a range of $1.51 (Batterham, 2014) to $33 (Richiardi, Pivotta, & Merletti, 2012). Given the significant range of costs, we offer the following advice on how to keep Facebook advertising costs to a minimum.

Advertisement type. Depending on the type of advert, the costs of advertising on Facebook vary considerably. Directing users to internal Facebook pages is generally less expensive than directing users to external Web sites, because the social network has an interest in keeping users within the platform. Furthermore, promoting a Facebook page, rather than an external one, makes it easier for participants to share the study with their friends. Batterham (2014) was able to reduce the cost-per-participant from $9.82 to $1.51 by redirecting users to a Facebook page that contained a link to the external online survey, rather than promoting the external link directly.

Bidding. Facebook asks advertisers to “bid” or set the maximum price that they are willing to pay for each click or each impression. This value is used by a bidding mechanism that selects adverts to be published to users; adverts accompanied by higher bids are more likely to be published. Facebook suggests a bid range guaranteeing an effective campaign in a given targeted group. Bidding below that suggested range still usually results in adverts being published, but it may take longer for an ad campaign to be completed. In our experience, it is possible to receive a desired number of clicks with half of the lowest recommended bid value. Besides, if the recruitment speed proves to be unsatisfactory, the bid can always be increased at a later time.

Target group. The cost-per-click (and therefore cost-per-participant) varies substantially across targeting criteria such as geographical location, demographic traits, or behaviors and preferences. For example, the suggested bid per click for women with a doctorate degree located in the United States was $0.43, compared to $0.35 in the United Kingdom. Thus, whenever feasible, researchers may consider recruiting participants from populations that are cheaper to reach.

Tracking Participants Across Studies

Another Facebook feature that can prove useful in recruitment is the ability to use a Facebook identification number to track participants across time and studies. This feature could also be used to protect privacy, as researchers can efficiently recognize participants without having to store any personally identifiable information.

Participant tracking could also reduce errors and help save participants’ time by minimizing the need to frequently re-enter the same information (e.g., basic demographic facts) in related studies. Additionally, participant tracking turns repeated submissions—a challenge in traditional Web-based studies, as noted by Birnbaum, 2004—into an opportunity for collecting longitudinal data. For example, myPersonality’s implementation of the IPIP proxy for NEO-PI-R was retaken over 1.15 million times, and over 5,000 participants answered this questionnaire more than 10 times. This kind of retest information can be used in longitudinal studies to examine substantive changes (e.g., in personality), and to evaluate the psychometric properties of the questionnaire, such as its test–retest reliability or the degree of misrepresentation.

Staying Connected: Building an Ex-Participant Community

Facebook features can be used to remain in contact with ex-participants, enabling researchers to build useful communities. For instance, users who “liked” the Facebook
Collecting Facebook Profile Data

The previous section focused on the benefits that Facebook offers in the context of recruiting and communicating with participants. Here we introduce another strength of this environment: access to an extraordinary amount of information about individuals and groups, recorded in a natural environment.

Facebook profile information includes self-reported information (e.g., schools attended, current workplace, age or gender), traces of behavior (e.g., status updates or likes), and data contributed by the others (e.g., photo tags or comments on a user’s wall). These data can be recorded retrospectively and thus often overcome the shortcomings of participants’ memory and biases during their participation in the study. For example, even the most motivated participants are unlikely to have enough time, attention, and knowledge to reliably report on past events attended (e.g., used in Han et al., 2012), the natural language used in their day-to-day conversations (e.g., used in Schwartz et al., 2013), or the shape of their own egocentric network. The latter often encompasses hundreds of agents and thousands of friend connections (e.g., used in Aral & Walker, 2012; Arnaboldi, Guazzini, & Passarella, 2013).

The full list of variables available to researchers through Facebook constantly changes as the platform expands and modifies its policies. At the time of writing, the following categories of information can be recorded with users’ permission:

1. Demographic profile, comprising a unique user ID, full name, profile picture, age, gender, relationship status, romantic interests, geographical location, place of origin, work and education history, biography, link to personal Web site, time zone, political and religious views, general interests, and lists of favorite music, movies, TV shows, books, quotes, and sports.

2. User-generated content consisting of status updates, photos, videos, comments on other people’s content or pages, links, and notes published by users or their friends. Each piece of content also contains metadata, such as the positions of people present in the picture, date of publication, list of people who liked it, its privacy settings, and so forth.

3. Social network structure containing a list of friends and the type of users’ connections. Connection types include friendships, family links (e.g., spouse, siblings, parents, or children), and followers.

4. User preferences and activities comprising their likes, group memberships, attended events, installed applications, and tags in photos or posts.

5. Information about users’ friends, such as demographic details and friends’ activities that are visible to a given user.

---

7 In 2012, each post on a Facebook page was viewed by 16% of the page’s fans, but in 2014, this has reduced to 6.5%. This is partly due to an increase in the number of Facebook pages and partly due to Facebook prioritizing paid newsfeed promotions. See http://techcrunch.com/2014/04/03/the-filtered-feed-problem.

8 A one-way connection, similar to following on Twitter, where the follower receives updates about the followed person, but not the other way round.

9 Likes can be used by Facebook users to endorse content such as status updates, comments, photos, links shared by friends, advertisements, Facebook pages, or external Web sites. Endorsements also result in users receiving updates on a given piece of content, such as comments on a liked status update or news published by a liked Facebook page. Likes were introduced on February 9, 2009.

10 Users can label or “tag” each other in photos and posts. Users can review/remove tags related to them (i.e., “untag”).

---
Pitfalls of Using Facebook Profile Data

Recording Facebook profile data offers an unprecedented opportunity to observe people in realistic environments. However, a major challenge in studying observational data, such as Facebook profile data, is to draw conclusions that are acceptably free from influences by overt biases, and to assess the impact of potential hidden biases. Several major issues have to be considered before drawing conclusions from Facebook profile data.

First, some parts of the Facebook profile are self-reported and other parts (e.g., behavioral traces) can be selectively removed by the profile owner. Thus, the quality of Facebook profile data may be affected by user-induced biases typical for self-reports, such as social desirability and intentional misrepresentation. Empirical evidence, however, indicates that Facebook profiles contain valid information about their creators. For example, Back et al. (2010) showed that Facebook profiles reflect the actual and not self-idealized personality profile of their owners. Also, Kosinski et al. (2013) showed that other elements of Facebook (such as gender, age, political views, religion, and Facebook likes) show consistent and meaningful relationships within Facebook profiles. Finally, in a subset of myPersonality participants (n = 28,628), where both the Facebook profile and self-reported information about gender were available, they matched in 98.8% of the cases. A somewhat lower accuracy (95%; n = 18,321) was recorded for age.11 Those results suggest that biases typical for self-reports are not particularly strong in this environment. This could stem from the fact that Facebook profiles are generated outside of a research context. Also, as Facebook friendships are usually preceded by real-world interactions (Lampe, Ellison, & Steinfield, 2006; Ross et al., 2009), inaccurate or enhanced profile invalid information may be difficult to maintain in a network of friends who can challenge false assertions (Lampe, Ellison, & Steinfield, 2007; Pempek, Yermolayeva, & Calvert, 2009).

Second, Facebook profile data is, to some extent, affected by the mechanics of the platform. The user experience on Facebook is highly personalized. This personalization includes the selection of news feed stories, advertisements, and even friend suggestions. Algorithms regulating individual experience are constantly evolving.

6. Private messages between users, usually written and sent through the instant messenger feature.

Facebook collects other forms of data in addition to those listed above, such as users’ visits on others’ profiles and the time that the user signed up to Facebook. However, at present, it does not share this information with users or third parties. The full list of accessible variables is available at http://developers.facebook.com/docs/graph-api/reference.

Interestingly, Facebook data offer insights reaching well beyond the boundaries of the platform. Liking content, posting photos, and interacting with friends represent behavioral patterns that are unlikely to be limited to Facebook. Adding a Facebook friend is, in most cases, merely a shadow of an interaction in the real world. Similarly, liking a Web site, musician, or activity is usually a proxy for behavior in other environments, such as visiting Web sites, listening to songs, or engaging in those activities. Moreover, Facebook stores data explicitly reaching into other online and offline environments. Examples include lists of real-world events attended, check-ins at geographical locations, and the likes generated by clicking a button located on an external Web site or by scanning a Quick Response (or QR) code.

Data collected from Facebook profiles can be further enriched using external models and services. For example, Pennebaker’s Linguistic Inquiry and Word Count (Pennebaker, Francis, & Booth, 2001) can be used to tag status updates with their emotional valence, and www.applymagicsauce.com (Kosinski, Stillwell, & Graepel, 2013) can translate Facebook likes into estimates of personality, intelligence, satisfaction with life, gender, age, political and religious views, and a number of other traits. Such procedures require skills and effort, but they can save participants’ time and reveal information that is difficult to access otherwise.

11 Several of the questionnaires published by myPersonality included questions related to participants’ age and gender. Neither of these sources can be treated as a “ground-truth” data, but potential inconsistencies provide a lower-bound estimate of the invalidity: Inconsistent data are inevitably invalid. Note that providing matching yet incorrect values is unlikely and pointless, given that those who did not want to reveal their real age or gender could simply skip relevant questions or hide the information from their Facebook profiles.
and, most likely, function differently for different users. As users are more likely to interact with content and people suggested to them by Facebook, their behavior is driven not only by their intrinsic goals and motivations, but also (to some unknown extent) by the Facebook algorithms constantly adjusting their exposure to content and friends. For example, friends’ photos that appear on a given user’s Facebook news feed are clearly more likely to be liked. Essentially, largely unknown effects of personalization represent a general class of confounding variables characteristic for observational research and deserve further study. Results likely to be affected by Facebook algorithms should therefore be interpreted with caution.

Furthermore, profile entries such as job position or political views are recorded using open text input fields, which are often equipped with an autocompletion mechanism (in the form of a pop-up list) that suggests potential entry values based on the first few letters typed in. This may improve the quality of the data (e.g., by discouraging spelling errors), but it can also introduce biases. For example, a user might have intended to identify himself as being a social psychologist, but may settle on social scientist if Facebook suggests the latter based on the first few letters. Consequently, a mere change in an autocompletion mechanism could affect the relative frequencies of the entries, and be misinterpreted as a real-world phenomenon (i.e., a sudden decrease in people’s self-identification as social psychologists).

Also, Facebook itself has changed considerably over the years, driving changes in users’ behavior. Likes are now a staple of the platform, despite only being introduced in 2009. Originally, users would list their favorite “activities,” “interests,” and other preferences in free-text fields; after 2009, these entries were converted into Facebook likes. Another change concerns status updates, which were originally preceded with the user’s name (i.e., “Alice Miller is . . .”), encouraging updates written in the third person. Nowadays, status updates are free to contain any text. Additionally, users can use privacy settings to limit access to parts of their profiles, and the platformwide privacy rules are constantly evolving. Initially, users could see and interact with everyone in the same university network, whereas today the visibility and interactions are usually limited—by default—to users’ immediate circle of friends. As with changes in autocompletion mechanisms, there is a risk that evolving privacy rules of the platform could be misinterpreted as a real-world phenomenon. For example, one may conclude that a certain group of users are deciding to make more of their content private, when in fact the default visibility setting of the content may have been altered for that group by Facebook, as part of a batch-release of new features.

Another concern related to the quality of the Facebook profile data relates to fake profiles. While it is easy to create a fake profile on Facebook, and use it to participate in the research such profiles are relatively easy to detect (Yang et al., 2011). For example, real users accumulate their friends and likes over a long period of time, but fake profiles are likely to be filled with likes and friends added in a single burst of activity. Also, egocentric networks of real users are relatively dense (i.e., their friends tend to know each other) and clustered in terms of geography and institutions (e.g., schools or employers). Friends of the fake users, on the other hand, are unlikely to know each other or be collocated. It is also relatively unlikely that participants will make the effort to use a fake account to participate in a study, when it is possible to simply adjust Facebook privacy settings to prevent third parties, including researchers, from recording their Facebook profile data.

### Collecting Self-Reports From Facebook Users

Collecting self-reports from Facebook users is similar to collecting such data in other online environments. In fact, an existing online survey or questionnaire can be integrated with Facebook merely by adding a fragment of HTML code (“Log in with Facebook” button14), allowing researchers to identify participants by their Facebook user ID and, with their consent, record selected data from their Facebook profiles.

Using Facebook to collect self-reports offers a number of advantages. As discussed before, participants can be reliably identified and tracked across studies, without necessarily revealing any personal details apart from their Facebook user ID. Moreover, many of the typical questions (e.g., related to demographics) can be skipped entirely, given that data can be obtained directly from the Facebook profile or inferred from the targeting approach used to promote the link to the study. Such an approach saves time and effort for both participants and researchers, while also helping to retain participants’ attention and preventing typing errors or other unintentional mistakes. Even if the desired information is not directly available through the Facebook profile data, it can often be inferred. For example, myPersonality did not contain any questions related to views on environmental issues, but such information can be extracted from other variables, including participants’ likes or status updates (e.g., liking “Stop Global Warming” or posting about related topics). Similarly, rather than collecting participants’ responses to personality items such as “I make friends easily,” researchers could measure the size and growth of their Facebook friendship networks, analyze the emotional valence of friends’ comments on their activities, or count the number of times they were tagged in other people’s pictures (Kosinski, Bachrach, Kohli, Stillwell, & Graepel, 2014; Kosinski et al., 2013; Youyou, Kosinski, & Stillwell, 2015).

---

12 For example, due to experiments conducted by a Facebook administrator to improve user experience or investigate their behavior (referred to as A/B testing).

13 Researchers could also attempt to collaborate or discuss their work with internal Facebook research teams who may be better informed about the functioning of Facebook algorithms and could help to avoid misinterpretations.

14 See http://developers.facebook.com/docs/facebook-login.

15 See the section Targeting: Recruitment Through Facebook Advertising.
Finally, it is important to recognize that participants may have disabilities (e.g., vision impairment) and use a variety of devices when accessing the study. For example, a growing number of people access the Internet (and Facebook) using mobile devices. It is thus important to make sure that a study presents itself well on large as well as small screens, and that the text and buttons are large enough (or scalable) to accommodate participants with disabilities. Naturally, such a requirement applies to all research, not just studies mediated via the Internet. In fact, Internet studies, which do not require participants to physically come to a research venue, are often more accessible for persons with disabilities.

**Improving the Quality of the Self-Reports**

Facebook offers a number of advantages when collecting self-reported data, but it shares some of the limitations of online research. First, low barriers to access the study, the ease of responding, and instant feedback may encourage the participants to rush through the study without paying much attention (J. A. Johnson, 2005; Kurtz & Parrish, 2001). Second, researchers have little or no control over the circumstances in which the study is being accessed, so it is possible that some participants simultaneously engage in other activities. Third, the lack of face-to-face contact increases the psychological distance between the researcher and the participants, which may decrease participants’ feeling of accountability (Caspi & Gorsky, 2006; Gosling et al., 2004). Finally, because Facebook participants come from diverse backgrounds, they may misunderstand instructions or test questions due to linguistic or cultural differences (J. A. Johnson, 2004, 2005).

In our experience, nevertheless, Facebook samples produce self-reported data of very high quality. The reliability16 of the psychometric scales published on the myPersonality application was similar to that reported in their manuals or respective standardization samples. For example, the average reliability of our implementation of the 100-item version of the IPIP proxy for NEO-PI-R equaled 0.91 (Stillwell & Kosinski, 2015), compared with 0.89 reported for the standardization sample (Goldberg, 1999). The same effect was observed for the 300-item version of this questionnaire (0.84, compared with 0.80). Similarly good results were obtained for the discriminant validity17 (John & Benet-Martinez, 2000), which was similar to the one obtained in paper-and-pencil samples (0.16, compared with 0.20; John & Srivastava, 1999; Stillwell & Kosinski, 2015). High reliability and discriminant validity in the myPersonality sample are even more remarkable given its diversity and international character. Though standardization samples are usually composed of carefully selected and highly motivated participants, myPersonality users were not controlled in any way and for many of them, English was not their first language. Finally, the quality of the self-reported data collected by myPersonality is also supported by the quality of the insights they offer. For example, Youyou et al. (2015) used self-reported personality scores to build a model predicting personality from respondents’ likes. The resulting personality predictions were more accurate than those made by participants’ spouses and family members.18 The high predictive accuracy of the computer model could not have been achieved if the self-reports used in its training were not of high accuracy.

The high quality of the data collected by the myPersonality project does not guarantee that all other studies will be equally successful. However, the size of our sample, the wide range of questionnaires used, and the longevity of the study suggest that it was not an accident. The example of myPersonality shows that participants recruited on Facebook, working in an uncontrolled environment and motivated solely by obtaining feedback on their scores, can provide researchers with high-quality data. Below we present several of the factors that are often undervalued by social scientists but, in our opinion, were key to obtaining high-quality data in our projects.

**Choosing the appropriate incentives.** In our experience, providing feedback on the scores or performance is one of the most efficient ways of compensating participants for their time and effort. Participants motivated by financial rewards or credit points—as are commonly studied undergraduate Psychology students or AMT workers—might be inattentive, respond dishonestly, act unnaturally, or perform half-heartedly. Such behavior invalidates not only the results of the study, but also the feedback. Therefore, rewarding them exclusively with feedback promotes attentive and honest participation, improving the validity of the results.

Some studies do not produce interesting feedback, but researchers can include elements that do so. Such an approach not only provides motivation for the participants, but also allows researchers to collect additional, potentially useful data. For example, while developing a new general intelligence test, researchers can mix new test items with ones from an established measure. Consequently, they can reward participants with the feedback on the scores and collect valuable cross-validation data.19

Another approach to motivating participants was tested in the case of the longest measure implemented on myPersonality: a 300 item-long personality inventory that required up to 1 hr of a participant’s attention. To get access to the questionnaire, participants were required to...
either fill out a series of uninteresting marketing surveys or make a donation ($4 USD) toward the cost of project hosting. This step severely limited (and presumably biased) the number of participants in this subpopulation, but the resulting data had excellent psychometric properties (Stillwell & Kosinski, 2015).

In general, we believe that financial incentives—although popular among social scientists and likely to boost the sample size (Doody & Sigurdson, 2003)—have numerous disadvantages. First, they do not reward people for responding honestly or behaving naturally, but merely for participating in the study. This reward structure can be detrimental to the quality of the data, especially considering that self-reports can easily be filled with dishonest or random information. Second, financial incentives are not scalable and become prohibitively expensive if the required sample size grows. Third, financial incentives can attract semiprofessional participants, who may have previous knowledge of the experimental design or materials used, and who have most likely participated in similar studies before (e.g., samples recruited through AMT or psychology students; Chandler et al., 2014). It is often easier to use the grant money to pay the participants than to design an engaging study, but we believe that financial incentives should be avoided whenever possible, due to their potentially detrimental effect on data quality.

Avoiding any coercion. When designing studies, researchers are too often completely preoccupied with their goals, and tend to forget about the participant’s experience. Consequently, participants are often forced to follow strict guidelines, having little control over the procedure and being presented with tedious and uninformative tasks. This approach may work relatively well in controlled laboratory settings, but online studies offer significantly less direct control over participants and their environment. Consequently, some of the established research practices may have to be dropped or redesigned.

One research practice that is particularly difficult to enforce in an online environment is limiting access to the study to only the chosen population (e.g., women). There are many ways in which researchers can try to limit access to a study, for example by rejecting participants that fail a prescreening question about their gender. However, we believe that such an approach is counterproductive. Some especially motivated members of other groups (in this case, males) could lie or use a fake Facebook account, creating an unnecessary source of invalid data that may be difficult to filter out. In practice, it is usually easier to remove participants that do not meet the target criteria before the analysis, rather than trying to filter them out at the data-collection stage.

Another practice that is likely to trigger dishonest responses or behaviors is when the researcher prevents participants from skipping questions or tasks. Even a small number of such invalid records can have a detrimental effect on the results of the study, so it is advisable to allow the participants to skip tasks. Afterward, incomplete protocols may be removed or the missing data imputed (Schafer & Graham, 2002). Additionally, we suggest that personal data or consent to record profile information should be requested at the end, rather than at the beginning of a study. Having acquainted themselves with the design of the study and the feedback it offers, participants might be more willing to provide honest responses and to trust researchers with access to their profiles.

Keeping it short and minimizing number of questions. The length of the study is another crucial issue to be considered when conducting online research. Decreased direct control over the participants increases the risk of them providing invalid responses or abandoning the study altogether. Thus, we believe that participants should be engaged gradually, for example by distributing incentives across the study. In the case of myPersonality, respondents could choose to receive their reward (basic feedback on their personality) after answering as few as 20 out of 100 personality questions. This ensured that the immediate barrier to participation was low and encouraged many of the participants to answer more questions or proceed to other questionnaires. Even if such an approach has several disadvantages (e.g., 20 questions did not provide enough information for many of the analyses, and early feedback could affect respondents’ further answers), we believe that it is better than encouraging participants to submit random responses or abandon the study prematurely.

Ethical Considerations

In the previous sections, we have argued that Facebook and other online environments have made it easier than ever before to observe individuals. Although this has the potential to greatly boost social science research, it also introduces new ethical challenges. The repercussions of misconduct in online human subjects research (HSR) could be far greater than ever imagined (Barchard & Williams, 2008; Hall & Flynn, 2001; Molokken-Ostvold, 2005). This was recently illustrated by the public reaction to researchers’ manipulation of users’ Facebook news feeds to study emotional contagion (Kramer, Guillory, & Hancock, 2014). Nevertheless, the protocols and guidelines related to designing online studies, storing data, and analyzing results are scarce and often contradictory (Solberg, 2010; Wilson et al., 2012). At the time of writing, the American Psychological Association’s Web site lists only three documents (Frankel & Siang, 1999; Hewson, 2003; Kraut et al., 2003) containing guidelines related to research on the Internet, the most recent being from 2003 (nearly 2 years before Facebook was founded). In the United Kingdom, the most recent edition of the Economic and Social Research Council Framework for Research Ethics (2012) simply states that Internet research presents new ethical dilemmas, without suggesting any guidelines beyond the requirement that any research conducted on the Internet should be...
subject to a full review by the appropriate Institutional Review Board (IRB).

The lack of clear guidelines is exacerbated by ever-accelerating technological progress; both researchers and IRB members may over or underestimate the threats to participants, thereby hindering benign projects or approving malignant ones. Both factors discourage social scientists from carrying out online research or submitting studies for review (Hall & Flynn, 2001; Singer & Vinson, 2001). As a result, an increasing fraction of HSR is carried out by computer scientists, who are often unconcerned about or unfamiliar with the ethical and social implications of HSR (Buchanan, Aycock, Dexter, Dittrich, & Hvizdak, 2011; Buchanan & Ess, 2009). Additionally, it discourages non-academic researchers (e.g., in commercial companies) from seeking any kind of ethical clearance or sharing their results with other scientists and the general public.

This trend is disconcerting, and not only because Facebook constitutes a powerful research tool and an important area of interest for social sciences. Thus, we hope to encourage the relevant bodies like IRBs, federal agencies like U.S. Department of Health and Human Services, and the American Psychological Association Ethics Committee to increase their focus on new research tools and environments, including Facebook.

Next, we discuss several aspects of consent and data collection that are, in our experience, especially relevant in the Facebook environment.

**Participants’ Control Over Data**

Facebook and other online environments have made it easier than ever before to share private information, in both greater detail and larger volume. Content that has traditionally been considered intensely private, and worthy of staunch legal protection, is now—at the click of a button—openly broadcast to one’s network of friends. Moreover, participants in Facebook-based studies seem to feel comfortable sharing those extensive records with researchers (Dwyer, Hiltz, & Passerini, 2007). For example, over 30% of myPersonality participants decided to volunteer the contents of their Facebook profiles, together with their personality, intelligence, and other psychometric scores (Stillwell & Kosinski, 2015).

Facebook users’ increased openness to share their personal information with researchers may be driven by several factors. First, Facebook users seem to have far greater control over their data than is usually assumed. Facebook profile data are shared with and scrutinized by hundreds of their friends and acquaintances. Information that a given user considers to be overly intimate, or that which casts him/her in a bad light, is unlikely to remain on the profile and thus would be unavailable to the researcher. Second, Facebook’s privacy settings enable users to revoke or limit access to data after it has been granted. Third, Facebook provides an additional layer of privacy by requiring users to confirm the consent given to applications (see Figure 1). Finally, traditional studies frequently involve personal interaction with the researcher in an unnatural laboratory setting. This can often make respondents anxious and self-aware, decreasing the chances of them being honest and behaving naturally (Smyth & Pearson, 2011). By contrast, studies based on Facebook and other online environments usually enable people to participate in a time and place of their own choice, yielding favorable effects on the generalizability and external validity of the results (Reips, 2000).

Facebook offers participants a relatively high degree of control over their data, but it is the researcher’s responsibility to weigh the costs and benefits of collecting and using personal user information (and to defer to an IRB in cases of doubt). The mere availability of data and participants’ willingness to share them does not grant researchers an automatic right to record and use them freely. The boundaries of processing the data should always be agreed on with the participant; this concern is especially acute in the context of findings showing that even a seemingly innocuous digital footprint, such as Facebook likes, can be used to infer highly intimate details, such as intelligence, personality, political views, and sexual orientation (Kosinski et al., 2014; Kosinski et al., 2013; Youyou et al., 2015). Fortunately, the Facebook environment enhances researchers’ ability to treat their participants as collaborators in research, rather than subjects. This could be achieved by ascertaining that the participants

---

**Figure 1**

An Example of a Facebook Consent Dialog Box, Shown After Offering Access to One’s Facebook Data to a Third Party

---

**Note.** See the online article for the color version of this figure.
obtain clear details about what information is going to be extracted from their data, and who will be able to access it;
- obtain feedback on their results or performance (where possible), regardless of whether they have consented to their data being stored;
- can review and retract any piece of information that researchers have collected based on their consent; and
- can request to receive a notification about the publication of any results based on their data. (We believe that, whenever possible, researchers should publish in open-access journals, and also endeavor to write and share nontechnical reports of results, such as blogs or articles in popular press.)

Granting participants the aforementioned rights would not only minimize the chances of researchers attempting to use data in an unacceptable way, but might also secure participants’ well-being, engagement, and trust.

**Consent Forms**

Consent forms used in social sciences are aimed at protecting participants and researchers. However, in the online environment, already replete with license and consent forms, users have been conditioned to click “Agree” without first reading the fine print (Böhme & Köpssel, 2010). This puts both the participants and the researchers at risk. Thus, researchers should take advantage of the design flexibility offered by the online environment to craft consent forms in a participant-friendly way.

The most important information could be clearly laid out in few sentences (or bullet points), using a large accessible font, with hyperlinks or pop-up dialog boxes enabling interested participants to obtain more details on each of the points. Such an approach could result in participants paying more attention to what they consent to and making a truly informed decision as to whether they want to participate or not. We hope that researchers and relevant institutions—including IRBs, federal agencies, and the American Psychological Association Ethics Committee—will investigate the feasibility of more participant-friendly formats of informed consents for online usage.

Importantly, as mentioned in the Participants’ Control Over Data section, we believe that people should be allowed to participate in the study (e.g., take the questionnaire or play a game) even if they declined to share their data or results with the researchers. Note that in some cases, consent could be requested (or rerequested) at the end of the participation, or along with the feedback; participants might be more open to sharing their data and results after experiencing the study.

**Boundaries of Individual Consent**

Perhaps the most challenging aspect of recording Facebook data relates to the vague boundaries between information that belongs solely to participants, and that which can be accessed with their consent, but belongs to or relates to other individuals. Examples of such data include pictures or videos featuring nonparticipants, messages or comments received from other people, or information about participant’s friends. These can all, at present, be recorded on Facebook with users’ consent.

We propose that it should be generally acceptable to use data generated by or containing references to nonparticipants, provided that the analyses are aimed *exclusively* at those directly participating in the study. For example, nonparticipants’ demographic profiles and network connections (that can be observed with participants’ consent) could be used to establish the parameters of participants’ egocentric social networks, or the gender ratio among their friends. Similarly, nonparticipants’ activity, such as photo tags, comments, or likes, could be analyzed in an aggregated form, to extract knowledge about participants, such as their popularity or social activity.

**Collecting Public Facebook Profile Data**

Some basic profile information is publicly available and even indexed by external search engines. Social scientists disagree on whether collecting publicly available data falls within the regulatory definition of HSR (Economic and Social Research Council, 2012), and whether it therefore requires participants’ consent and IRB approval (Schultze & Mason, 2012; Solberg, 2010; Wilson et al., 2012). Some scholars point out that the border between public and private is not determined by accessibility, but by social norms and practices (Frankel & Siang, 1999; Waskul, 1996). This point is illustrated by the example of a small town in which everyone knows intimate details about everyone else, but people pretend not to know those facts that would be considered personal (Schultze & Mason, 2012). Others argue that mining public data is an equivalent to conducting archival research, a method frequently employed in disciplines such as history, art criticism, and literature, which rarely involve rules for the protection of human subjects (Bruckman, 2002; Herring, 1996).

We agree with the latter argument and propose that using public Facebook profile data (e.g., data that are publicly available to any Facebook user and not only to one’s Facebook friends) should not require participants’ consent if the following conditions are met:

- It is reasonable to assume that the data were knowingly made public by the individuals
- Data are anonymized after collection and no attempts are made to de-anonymize them
- There is no interaction or communication with the individuals in the sample
- No information that can be attributed to a single individual, including demographic profiles and samples of text or other content, is to be published or used to illustrate the results of the study.

If any one of the above conditions is not met, the study should be closely scrutinized by an IRB.
Discuss Ethical Considerations and Privacy Consequences When Publishing Results

Finally, following Schultze and Mason (2012), we encourage researchers conducting HSR on Facebook (and other online environments) to include a discussion of ethical considerations related to the design of a given study in their publications. Furthermore, we believe that the study should discuss the ethical implications of the findings. Such an approach would ensure that the authors have properly considered the ethical aspects of their own work, and would support the evolution of standards and norms in the quickly changing technological environment.

Conclusions

A growing proportion of human activities, such as social interactions, entertainment, shopping, and information-gathering, are now mediated by digital devices and services such as Facebook. Digitally mediated behaviors such as these can easily be recorded and analyzed, fueling the emergence of computational social science (Lazer et al., 2009; Markowitz, Blaszczewicz, Montag, Switala, & Schlaepfer, 2014). They also facilitate the transition from small-scale experiments and observational studies to large-scale projects based on thousands or millions of individuals (Backstrom, Boldi, Rosa, Ugander, & Vigna, 2012; Eichstaedt et al., 2015; Kosinski et al., 2014; Kosinski et al., 2013; Kramer et al., 2014; Ugander, Karrer, Backstrom, & Marlow, 2011; Youyou et al., 2015). Observing or experimenting with large samples enables scientists to minimize the problem of sampling error typical to social science and to detect patterns that might not be apparent in smaller samples. It also offers unprecedented insights into the dynamics and organization of individual behavior and social systems, with the potential to radically improve our understanding of human psychology (Lazer et al., 2009). The same processes and technologies driving the emergence of computational social science are also rapidly transforming the human environment. This drives the need to reexamine the relevance of established social science theories, and use modern technology to develop new ones.

However, researching psychological phenomena in the digital environment requires skills that are relatively uncommon among social scientists, such as recording, storing, processing, and analyzing large databases. For example, the combined size of all databases published on the myPersonality project Web site exceeds 50GB, and some of the tables contain hundreds of millions of rows. Processing or analyzing tables of this size and complexity requires at least a basic knowledge of scientific programming (e.g., R or Python) and database management (such as MySQL). Additionally, it is necessary to truly understand the online environment under investigation. A degree of familiarity can be achieved by studying the literature, but this is no substitute for personal experience with the platform or, in other words, becoming an active Facebook user.

Another example of a potential drawback is that Facebook allows users to list their family members, but some users choose instead to list their close friends as brothers or sisters rather than, or in addition to, their actual family members. Researchers who are not active Facebook users are prone to invalidating their results by overlooking or misunderstanding peculiarities such as this.

As social scientists are relatively slow in embracing the necessary practical skills (Lazer et al., 2009), data-driven human subjects research is increasingly ceded to computer scientists and engineers, who often lack the appropriate theoretical background and ethical standards (Buchanan et al., 2011; Hall & Flynn, 2001). We strongly encourage our fellow social scientists to not only train themselves in modern computation methods, but to also immerse themselves in new human environments, including Facebook. These digital platforms are more than an object of study. Being as rich and diverse as the complex human environments from which they emerge, digital platforms offer new opportunities for social science research and new challenges for researchers in their own right. With proper training, traditional social science studies can be conducted online at a lower cost and larger scale than ever before.

REFERENCES


Richardi, L., Pivotte, E., & Merletti, F. (2012). Recruiting study participants through Facebook. Epidemiology, 23, 175. http://dx.doi.org/10.1097/EDE.0b013e31823b5ec4


