Invited Forum: Challenges in Making Data Available
General Article

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(1) 86–94

DOI: 10.1177/2515245917744281
www.psychologicalscience.org/AMPPS

Open Sharing of Data on Close Relationships and Other Sensitive Social Psychological Topics: Challenges, Tools, and Future Directions

Samantha Joel[1], Paul W. Eastwick[2], and Eli J. Finkel[3,4]
[1]Department of Psychology, University of Utah; [2]Department of Psychology, University of California, Davis;
[3]Department of Psychology, Northwestern University; and [4]Kellogg School of Management, Northwestern University

Abstract
This article reports on an adversarial (but friendly) collaboration examining the issues that lie at the intersection of confidentiality and open-data practices. We describe the process we followed to share our data for a speed-dating article we recently published in *Psychological Science* (Joel, Eastwick, & Finkel, 2017) and provide a summary of the issues we considered and addressed along the way. As we drafted the present article, the third author became unsure, in retrospect, about some of the procedures we had followed, especially if our approach were to be perceived as a model for open-data decisions in other, more typical cases involving nonindependent data. This article addresses these concerns, but also identifies areas of consensus. All three authors agree that there remains an unmet need for guidelines and other resources to help researchers address the challenges of sharing data that cover sensitive topics, particularly nonindependent data collected from pairs and groups (e.g., romantic couples, work teams, therapy groups). We conclude with a discussion of new tools that could be developed to help scholars who have collected such data to increase the transparency of their research while simultaneously protecting the confidentiality of the participants.

Keywords
relationships, marriage, mate selection, open data, nonindependent data, sensitive data

Received 8/19/17; Revision accepted 10/31/17

Psychological scientists face a pressing need to improve the transparency of their research. Within the past decade, new evidence has suggested that the reproducibility of psychological findings has considerable room for improvement (e.g., Munafò et al., 2017; Open Science Collaboration, 2015; Vazire, 2017). Openly sharing data improves the credibility of published findings because access to raw data allows scientists to confirm, critique, and improve upon each other's work (e.g., Asendorpf et al., 2013; Nosek, Spies, & Motyl, 2012; Vision, 2010). Findings that are paired with open data are more trustworthy because other researchers can use various procedures to evaluate the likelihood that the reported results did not rely on error (e.g., Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016) or biased analysis strategies (researcher degrees of freedom; Simmons, Nelson, & Simonsohn, 2011).

Sharing data also maximizes the contributions that a particular study can make to the literature because the data are available for reanalysis and meta-analytic aggregation over time (e.g., Wicherts, 2016); in many cases, scholars can use the data sets to conduct novel analyses that the original researcher or researchers would not have conducted. Sharing one's data is also a powerful educational tool. One primary goal of the classic quantitative journal article is to inform other researchers about the fruits of one's labor, typically using summary statistics. Opening these data up to colleagues makes this educational process far more comprehensive—and

Corresponding Author:
Samantha Joel, University of Utah, Psychology Department, 380 South 1530 East, BEH S Room 811b, Salt Lake City, UT 84112
E-mail: samantha.joel@psych.utah.edu

ultimately more persuasive—because other scholars can experience for themselves the pathway from raw data to published conclusions.

At the same time, the discipline of psychology encompasses many sensitive, personal topics for which confidentiality is an important concern. In social psychological research, for example, participants may disclose their private religious views, political beliefs, feelings of prejudice toward members of other groups, personal insecurities, hurtful experiences, opinions about people with whom they have close relationships, and willingness to harm other people (see Gilovich, Keltner, Chen, & Nisbett, 2016). When participating in research studies on such topics, participants place a great deal of trust in the researchers to protect their confidentiality; indeed, if confidentiality were not guaranteed, it is unclear whether participants' answers to sensitive questions would be truthful and, thus, worth studying in the first place. Breaches in confidentiality are unacceptable: Not only do they risk meaningful negative consequences to the participants (e.g., if the wrong information were to reach their family, friends, or employers), but they violate the researcher's contract with the participants and, in turn, could erode the public's trust in researchers and willingness to share their personal experiences for research purposes.

Protecting confidentiality is not always as straightforward as it might seem. In theory, one can de-identify a data set by simply removing personal information (e.g., names, e-mail addresses) before making the data publicly available. However, variables that do not appear to allow personal identification can sometimes be used, especially in combination with other data, to re-identify a data set (e.g., Samarati, 2001). For example, psychology studies often include age and ethnicity in analyses, and manuscripts often report the host university and approximate year during which a particular study was conducted. If the data from such a study were made publicly available, the researcher would have an ethical imperative to ensure that the confidentiality of particularly identifiable students (e.g., the only 34-year-old student who self-identified as a Pacific Islander and was a freshman at University X in 2013) was not inadvertently being compromised.

Special issues arise regarding the sharing of nonindependent data: data for which observations are nested within groups, such as couples or families. When the data are independent, one must ensure only that an outside observer cannot identify the responses of particular participants. However, it is uniquely challenging to de-identify nonindependent data in a way that protects participants from having their data identified by another participant (Finkel, Eastwick, & Reis, 2015). For example, in a study in which married partners separately disclose their true feelings about their marriages, the data must not be shared in a way that would make it possible for partners to use their insider knowledge about the study (e.g., their own responses to the questions) to locate each other's responses. It is plausible that some will go looking for their partners' responses, given that many romantic partners are indeed motivated to snoop into each other's private information (e.g., Vinkers, Finkenauer, & Hawk, 2010). Thus, the onus is on the researcher to ensure that participants cannot discover via a "confidential" study that their partners do not love them, or are no longer attracted to them, or still pine for an ex.

Overall, open sharing of data offers a powerful way to increase the reproducibility and replicability of research findings, as well as the overall contribution of the data to the scientific community. Yet, if researchers in fields that rely on data that cover sensitive topics (e.g., relationship science) wish to reap these benefits, they must develop and adhere to procedures for sharing data that also protect participants' confidentiality. This article reports on a set of issues and complications that can arise when scholars seek to publicly post data on sensitive social psychological topics, especially nonindependent data from dyads or other groups. (We do not address the sharing of other kinds of sensitive data, such as medical data or data from vulnerable or stigmatized groups, in large part because of the robust literatures in those areas.) We offer a detailed discussion of how we were able to circumvent these complications in a recent case that concluded with a data set made openly available to researchers.

Our Own Foray Into Sharing Nonindependent Data

We recently concluded our first attempt at openly sharing sensitive, nonindependent data. Specifically, we sought to share the two speed-dating data sets required to reproduce the findings in our recent *Psychological Science* article (Joel, Eastwick, & Finkel, 2017). In these studies, conducted in 2005 and 2007, a total of 350 participants completed a long intake questionnaire relevant to mate selection (i.e., participants' own traits, ideal-partner preferences, and other individual difference measures). Each participant then attended a heterosexual speed-dating event, which included approximately 12 men and 12 women. At the event, each participant had a 4-min speed date with each opposite-sex participant (total $N = 2{,}050$ speed dates) and then completed a questionnaire on those speed dates. Our empirical article used all relevant measures from the intake questionnaire to predict romantic desire via a machine-learning method called random forests.

Thus, to enable other researchers to reproduce our findings, we needed to share data from more than 100 self-report measures collected from each sample.

As elaborated in this section, we consulted various sources and considered several options before choosing the UK Data Service (www.ukdataservice.ac.uk) as our data repository. Data uploaded to this repository are stored in a time-stamped, noneditable, and nonretractable format. We used the UK Data Service's "safeguarded access" option, which requires researchers to register an account on the Web site (free of charge) before accessing the data. Registration requires a stated affiliation with a suitable professional organization (e.g., a university) and a valid institutional e-mail address. The user must also agree to an End User License, which stipulates that data must be kept confidential.

As we sought to share our nonindependent data, we asked ourselves three key questions: (a) Is anonymization possible? (b) Did the consent process address data sharing? and (c) How great is the potential for harm? We weighed our responses to these questions against the benefits of data sharing (discussed in the introduction) and concluded that openly sharing the data was the correct decision in this case. In the process of writing the present article, however, the consensus that had characterized our efforts began to crack; thus, this section of the article can be viewed as an (amicable) adversarial collaboration. In particular, Eli developed some concerns about whether the procedures we had used for assessing risks to our participants were optimal, especially if the present article were to be perceived as a prescriptive guide for the procedures that other researchers should follow vis-à-vis their own data sets. We describe the process that we used before turning to Eli's retrospective reservations about it. Paul and Sam then offer rejoinders to Eli's concerns. We conclude this article with a consensus section oriented toward maximizing the public sharing of data while minimizing risks to participants.

The process we used

In this subsection, we discuss the three primary questions we asked ourselves when evaluating whether it was acceptable to post our data and provide the answers we developed in response to these questions.

Is anonymization possible? We first made every effort to de-identify the data. As these were nonindependent data, we carefully considered the identifier variables that linked people's responses with other participants with whom they interacted as part of the study (e.g., which speed-dating event they attended, which usernames correspond to their speed-dating partners, the order in which participants met their speed-dating partners). With access to these variables, people could conceivably use their knowledge of their own responses to identify others' responses. For example, participants could use their knowledge of their own characteristics and preferences (e.g., their prediction of what percentage of speed-dating partners they would like) to locate their own responses, and then use their own ID number to find out how specific other participants in the study perceived them (e.g., how attractive or desperate they seemed).

We took a number of steps to mitigate this concern, including removing all open-ended responses, removing personal identifiers other than gender (e.g., age, ethnicity, zip code, birthday), and removing uncentered responses to individual items whenever possible. Specifically, we removed any variables that indicated membership in an underrepresented group, as such variables can lead a person to be identifiable in the context of a research sample (e.g., if that person was one of the only participants of a certain age, ethnicity, etc.). We also took the unusual step of removing identifier variables from the dyadic data set: Although identifier variables are typically required to reproduce analyses that account for nonindependence (e.g., in multilevel models), our particular analytic strategy (machine learning) dealt with nonindependence prior to data analysis. That is, our analyses accounted for nonindependence through centering the dependent measure rather than by using person-level identifiers, and so the identifiers are not required to reproduce the analyses reported in the article. Our final materials still included gender and a large number of responses to individual items, as well as the years and geographical location of the speed-dating events. However, our best judgment was that (a) the likelihood of participants identifying their own responses from this combined information was small, and, more important, (b) the likelihood of participants identifying each other's responses was minuscule—particularly because identifiers linking participants' responses to the target being rated had been removed and because data collection had taken place 10 to 12 years earlier.

Had we been unable to remove the identifier variables, the other anonymization steps would have been even more crucial. To maintain confidentiality in a set of nonindependent data that contains identifier variables, one must anonymize the data to the point that the likelihood of a participant identifying his or her own responses would be miniscule. This could potentially be accomplished by removing not only demographic information, but also all individual items for which participants might conceivably draw on recollections of their own ratings to identify their own data in the data set. Ideally, the data set would be left with only centered or aggregate variables that have no

extreme outlier responses, so that it would be impossible for participants to discern which responses belonged to them (and, by extension, which responses belonged to their partners, friends, etc.). Alternatively, the data set could be shared in a repository that effectively prevents participants from accessing the raw values from the data set. We return to a discussion of repository options later in the article.

Did the consent process address data sharing? Participants may have a variety of preferences surrounding the sharing of their data, and these preferences may be either in favor of sharing (e.g., maximizing the contribution of their time and effort) or in opposition to sharing (e.g., concerns about confidentiality and sensitivity). Ideally, these preferences should be considered during the consent process (Cummings, Zagrodney, & Day, 2015). As our *Psychological Science* article relied on existing data sets—and consent was given in 2005 or 2007—we verified with the institutional review board (IRB) at Northwestern University that sharing these data sets on the UK Data Service was not in breach of the studies' IRB protocol. To our surprise, the IRB informed us that because the study-specific IRB protocol had been closed and the data were de-identified, the decision of whether to share data was no longer under the purview of the IRB (see also Burnham, 2014).

Thus, we revisited the language on the consent form signed by our participants a decade earlier, which read:

> Results of this study may be used for teaching, research, publications, or presentations at scientific meetings. If your individual results are discussed, your identity will be protected by using a study code number rather than your name or other identifying information.

Because the consent form had included this language about sharing (individual or aggregate) results for research and teaching purposes, our assessment was that we could (and should) share these data with the academic community through the UK Data Service's safeguarded-access option. If we had shared these data openly with all members of the public, it is not clear how we would have ensured that the data would be used only for scholarly purposes. In the future, in light of new open-science practices, we recommend stating more explicitly in the consent form that the data will be de-identified and shared (e.g., "Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public," as recommended by ICPSR, 2017).

How great is the potential for harm? Our third consideration was the degree of sensitivity of the data. What were the risks associated with linking people's identities to their responses to the specific measures in our studies? Our assessment was that these risks were reasonably low; although our studies included a number of somewhat delicate or embarrassing measures (e.g., desperation to find a romantic partner, other people's perceptions of participants' attractiveness), they did not include what we would deem to be measures with great potential for harm (e.g., infidelity or abuse, thoughts about divorce). Also, because the data were 10 to 12 years old at the time we considered sharing them, even if anonymity were breached (which we deemed extremely unlikely, as explained earlier), we viewed it as unlikely that any openly shared data would be harmful enough to meaningfully harm anyone's ongoing relationships. After carefully considering and discussing these issues, we concluded that the UK Data Service's safeguarded-access option sufficiently mitigated any remaining risk.

Retrospective wariness: Eli's squeamishness about relying on the subjective judgments of individual researchers concerning the appropriateness of sharing data

The conversation about data sharing has evolved in just a few short years: Straightforward exhortations to share data publicly have been replaced by nuanced discussions that detail the challenges and risks of doing so (e.g., Tackett et al., 2017). I (Eli)[1] was part of a team that had described these challenges in the past (Finkel et al., 2015), but at that time, I did not offer solutions that would allow for the sharing of nonindependent data. Thus, I was, and remain, proud of how hard we worked to break through the barriers to open practices for data like ours and how diligent we were about thinking through the complexities of doing so. But writing the present article—especially the parts discussing the questions we asked ourselves, and how we answered them—made me think about our decision process in a new way. Although I continue to believe, and hope, that we made the correct choices in how we openly shared our data, I am now less confident than I was about the process we used for making those choices. I became concerned that our approach—asking ourselves complex ethical, technological, and legal questions and trusting our best intuitions to answer them—might become a model for how scholars should approach these issues. The relevant issues are complex in every case, but much more so in the vast majority of cases involving nonindependent data than they were in our case (because, as noted previously, our machine-learning procedures allowed us to exclude a person-level identifier variable). Therefore, I am not confident

that our approach should serve as a model for other researchers. My hope is that going public with the internal debates that Sam, Paul, and I had will help other scholars grappling with these sorts of issues.

Regarding anonymization, we had concluded that the likelihood that participants will be able to identify their own responses in our data set is small, and that the likelihood that participants will be able to identify other participants' responses is minuscule. But data hacking and computer security are major industries, and a clear conclusion from experts in that area is that apparently secure data often are not secure (Zimmer, 2010). For example, if a participant recalls that on the speed-dating intake questionnaire, she said she expected that she would say "yes" to 85% of her speed dates, she could figure out that, say, only three of the people who gave that exact answer were women. And she could use other variables to figure out which row of data is hers. As I reflected more deeply on these issues, I began to wonder: By what criteria are psychological scientists (ourselves or other researchers) qualified to determine that participants could not identify their own data—and also the data of their partner or roommate or coworker?

Regarding consent, we had concluded "that we could (and should) share these data with the academic community through the UK Data Service's safeguarded-access option" (p. 89). But, upon reflection, the meaning of "the academic community" is ambiguous. What it actually means for the UK Data Service is anybody with an institutional (e.g., .edu) e-mail address who is willing to sign a confidentiality agreement. But consider how common it is for relationships researchers to study college couples, a population for whom institutional e-mail addresses are hardly rare. If the data-sharing approach we adopted for our *Psychological Science* article were to become normative, it would be trivially easy for participants in many studies to access their partner's raw data.

Regarding potential for harm, we had concluded that the risks are "reasonably low" in our case because "although our studies included a number of somewhat delicate or embarrassing measures (e.g., desperation to find a romantic partner, other people's perceptions of participants' attractiveness), they did not include what we would deem to be measures with great potential for harm (e.g., infidelity or abuse, thoughts about divorce)" (p. 89). But here again, it is not clear to me that researchers are qualified to make such judgments on behalf of participants. It seems plausible that some participants would view their response to our question asking how romantically desperate they felt to be sensitive. What criteria should psychological scientists use to determine whether data are too sensitive to share, especially when they might have strong a priori motivation toward or against making their data openly available?

Overall, although we worked hard to figure out a way to share our data for the *Psychological Science* article while respecting the rights of our participants, I now wonder whether we optimally weighed our eagerness to make our data open against the potential risks of doing so. Because our studies had some extremely rare features—especially that the data were more than a decade old and that we could eliminate identifier variables while still including all data required to reproduce our analyses—I continue to believe (and hope) that we probably got it right. But my sense is that our procedures—using our best intuition to answer self-interrogations—may be excessively risky for the vast majority of nonindependent data in psychological science.

Rejoinders (from Paul)

I agree with the core of Eli's critiques; I cannot say with perfect certainty that our data are completely anonymized, that we interpreted "the academic community" appropriately, or that there is zero potential for harm. But I wish to place our strategy—"using our best intuition to answer self-interrogations"—in a slightly broader context.

For example, here are two strategies I like less: (a) mindlessly and automatically posting nonindependent data for the general public to access and (b) never even considering posting nonindependent data for the general public to access. My coauthors and I found an optimal balance between these two extreme choices; we thought deeply about the potential risks of sharing nonindependent data, worked hard to address those risks, and posted the data at the end of this long process.

Here is a strategy I like more: asking a group of data-science professionals to evaluate whether we properly anonymized our data set. We did not do this, primarily because we do not know of such a service for academics. We were surprised to learn that IRBs typically do not evaluate de-identification procedures systematically; in fact, it is not clear that Certified IRB Professionals receive training in data anonymization that is more comprehensive than the limited training my coauthors and I have received (see also Meyer, 2018, this issue). Overall, I found it frustrating and disappointing that so much digital ink has been spilled over the importance of data sharing but very little of it has been devoted to helping researchers with (modestly) complex data sets join this brave new world.

So, in the absence of guidance, we tried our best to assess whether we were ethical in our data sharing. It

was not the optimal strategy to use our own intuitions to answer complex ethical, technological, and legal questions, I agree, but it beat both the mindless-sharing and the obstinate-refusal-to-share strategies.

Rejoinders (from Sam)

Eli raises many valuable critiques that social psychological researchers must consider as they move forward with open practices. I would like to offer two rejoinders concerning the risks associated with generalizing our procedure to other cases. First, although the data for our machine-learning article are indeed unusually safe in the ways that Eli mentions (e.g., the identifier variables have been removed), they also carry the unusual risk of including raw data for a variety of individual items. In more normative cases, reproducing analyses frequently requires only aggregated or centered variables. It would be considerably more difficult—if not impossible—for participants to identify themselves in a data set containing no raw values. If it is not possible for participants to use their knowledge of their own responses to identify themselves, then they cannot use an identifier variable to identify other participants, even if that identifier variable remains in the data set.

More broadly, Eli offers several examples of ways in which a single safeguard might fail. However, my view is that it is important to consider the use of these safeguards in combination. Assuming that the chance of each protection failing is independent, the risk of a confidentiality breach decreases exponentially with each new protection that is added. Researchers looking to share their data must consider the extent to which a given *combination* of safeguards protects confidentiality in the context of their particular study. For example, in the case of nonindependent data that have been de-identified and shared in a vetted repository, one must consider the combined odds that a given participant provided a unique outlier response that he or she remembers providing, is able to use this information to identify a confidential response from another participant, is able to meet the repository's vetting criteria by the time the data are shared, and is motivated and resourceful enough to find the repository where the data are stored. One may also wish to consider the odds of the participant identifying the data through more traditional means (e.g., physically stealing a less de-identified version of the data from the lab or hacking the researcher's computer) to determine what new risks are truly being introduced through open data sharing.

Overall, though, I agree with Eli's central argument that it is questionable how qualified any individual researcher is to decide what safeguards are sufficient to protect participants' confidentiality and under what circumstances. This brings the three of us to our strongest area of consensus: the discipline's need for clearer guidelines to help researchers navigate these complex issues.

New Tools We Would Like to See

Our experiences in pursuing a way to make our speed-dating data open and in writing the present article have led us to conclude that there is still a strong need for practical tools and guidelines to help researchers make their data open while also protecting the confidentiality of the participants. We now discuss several tools that we hope to see developed as the open-science movement continues to gain traction.

More, and better, data-repository options

When researchers cannot make their data fully open because of ethical considerations, the journal *Psychological Science* has indicated that it may be sufficient to use "a repository that vets requests for access to data" (Lindsay, 2017, p. 701). That is, in the interests of keeping data secure while also preventing researchers from ignoring requests to share their data, a third party can be made responsible for sharing the data with every qualified professional who requests access (and with only qualified professionals). As of April 2017, we found a dearth of data repositories that offered this service. SAGE's recommended search site, http://www.re3data.org/, yielded few options, as the vast majority of databases allow researchers to either make their data fully public or keep their data fully private, and do not provide the option of allowing access to the data to be vetted by a third party.

We received a variety of helpful suggestions regarding this issue on the online Facebook Group PsychMAP (Joel, 2017). We are particularly grateful to Debbie Hyden for suggesting the UK Data Service, which is the repository that we ultimately used. Two other promising suggestions we received were the Open Science Framework (OSF; https://osf.io) and the Harvard Dataverse (https://dataverse.harvard.edu); however, neither currently offers any level of third-party vetting. OSF also does not yet offer a way to register, or time-stamp, data while keeping it private indefinitely; embargos are automatically released after a maximum of 4 years. Another suggestion we received was ICPSR (https://www.icpsr.umich.edu/). Researchers must pay a fee of $350 and also obtain IRB approval before ICPSR will grant them access to the requested data. The data are then mailed to the researchers on a compact disc. Given these barriers to access, this repository seemed too restrictive for our purposes. However, ICPSR may offer

a useful compromise for highly sensitive studies (see our earlier discussion on potential for harm).

Sam also inquired about the issue of nonindependent data at a seminar on scientific integrity (Nelson, Simonsohn, & Simmons, 2017). The presenters had the innovative suggestion that a site such as Shiny Apps (https://www.shinyapps.io/) might be used to create an application that would allow researchers to run analyses on a particular data set without being able to access the raw data themselves. The availability of such a service could greatly help researchers with sensitive data to make their research more transparent and reproducible. Alternatively, one could generate a "mimicked" data set that reproduces the central features of the real data set (e.g., using the R package *synthpop*) and make it available to the public. Again, although this option is not as transparent as sharing the original data, it may offer a useful compromise for studies with particularly challenging confidentiality issues.

Overall, there appears to be a strong unmet need for services that allow people to share sensitive data both openly and safely: The UK Data Service was the only appropriate (albeit imperfect) repository we found for our data. One promising future avenue may be for university libraries themselves to host and vet access to data locally. Alternatively, professional organizations (e.g., Association for Psychological Science, or APS; American Psychological Association, or APA) might consider offering an application or vetting service in the future as a way to further encourage open data. Regardless of who vets requests for and grants access to a data set, considerable thought needs to go into the vetting criteria.

Guidelines, training, and other resources

At what point is a data set sufficiently de-identified that it can be made public? What role should the consent process play when a researcher is considering open-data options? What level of restriction is appropriate for what level of data sensitivity? Our discipline is still in need of clear, prescriptive guidelines that address these issues at the intersection of confidentiality and open-data practices, so that researchers are not relying so much on their own intuitions when making these decisions. We had to rely on our own intuitions, and it is possible that they were wrong. Data science is a vast field of inquiry, and if psychology is to be a mature, 21st-century discipline with respect to technological sophistication and transparency, researchers must tap into that expertise.

One route to achieving this would be for psychological scientists to develop their own training. Currently, it is rare for psychology graduate programs and statistics classes to include training on data anonymization, despite this being an increasingly necessary skill as psychology moves toward more transparent research practices, and despite the (frequently nonobvious) complexities that can be involved in proper anonymization. We advocate for more accessible resources and training on how to properly and thoroughly de-identify a data set—resources and training akin to those that are provided to health researchers (e.g., U.S. Department of Health and Human Services, 2015)—particularly in fields within psychology that frequently involve sensitive data.

Another option, especially in the near term and in the absence of established protocols within the discipline, would be to outsource the task of ensuring confidentiality to trained professionals. Universities or professional organizations, such as APS and APA, could establish a service whereby data scientists are available to double-check data sets with respect to sensitivity and anonymity before research teams share them.

Protocols to increase compliance with data-sharing requirements

Despite a researcher's best efforts, there may still be data sets that are too sensitive to share via a third party, particularly with currently available tools. However, there remains the option of sharing the data privately with qualified scientists who request them. In fact, the discipline of psychology has long required researchers to share their data with competent professionals who wish to verify the results (APA, 2002). However, this requirement is unenforced, and adherence to it is poor; for example, one team of researchers was able to obtain the original data for only 64 out of 249 studies published in 2004; in other words, 73% of authors failed to share their data (Wicherts, Borsboom, Kats, & Molenaar, 2006). Another possible solution to the problem of sensitive data, then, would be for professional organizations to implement protocols to better enforce data-sharing requirements. For example, APS and APA could keep a record of requests that researchers make for data published in their journals. These organizations could then follow up with researchers who fail to respond to requests for their data within a reasonable time frame, and impose consequences for repeated, unjustified noncompliance. It is worth nothing, however, that this approach requires potentially low-power researchers to directly request data from other, potentially high-power, researchers. An external application or vetting system is preferable whenever possible because it allows researchers to access data anonymously, eliminating any status and reputation concerns they might have.

Conclusions

Nonindependent and other kinds of sensitive data pose an important challenge for the open-science movement, and there remains a strong need for workable solutions. Some tools that we believe would facilitate the open sharing of sensitive data include (a) services, potentially hosted by universities or professional associations, that are able to vet or protect access to data; (b) accessible resources for and training on sharing sensitive data safely; and (c) protocols that improve compliance with data-sharing requests. We predict that the development of such tools would help fields that use sensitive data to benefit more fully from open-science practices.

Action Editor

Daniel J. Simons served as action editor for this article.

Author Contributions

E. J. Finkel and P. W. Eastwick generated the idea for this article. S. Joel wrote the first draft of the manuscript, and P. W. Eastwick and E. J. Finkel provided critical revisions and additions. All the authors approved the final submitted version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Note

1. This section and the next two sections present each author's individual views and are therefore written from a first-person perspective.

References

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*, 1060–1073.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119.

Burnham, B. (2014, February 5). Open data and IRBs [Web log post]. Retrieved from http://osc.centerforopenscience.org/2014/02/05/open-data-and-IRBs/

Cummings, J. A., Zagrodney, J. M., & Day, E. (2015). Impact of open data policies on consent to participate in human subjects research: Discrepancies between participant action and reported concerns. *PLOS ONE, 10*(5), Article 0125208. doi:10.1371/journal.pone.0125208

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology*, *108*, 275–297.

Gilovich, T., Keltner, D., Chen, S., & Nisbett, R. E. (2016). *Social psychology* (4th ed.). New York, NY: W. W. Norton.

ICPSR. (2017). *Recommended informed consent language for data sharing*. Retrieved from https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confientiality/conf-language.html

Joel, S. (2017, April 24). Data repositories that can vet requests [Online forum comment]. Retrieved from facebook.com/groups/psychmap

Joel, S., Eastwick, P. W., & Finkel, E. J. (2017). Is romantic desire predictable? Machine learning applied to initial romantic attraction. *Psychological Science*, *28*, 1478–1489.

Lindsay, D. S. (2017). Sharing data and materials in *Psychological Science*. *Psychological Science*, *28*, 699–702.

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science, 1*, 131–144.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), Article 0021. doi:10.1038/s41562-016-0021

Nelson, L., Simonsohn, U., & Simmons, J. (2017, April). *Scientific integrity*. Lecture presented at the University of Utah Travelling Scholar Seminar Series, Salt Lake City, UT.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1226.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. doi:10.1126/science.aac4716

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, *13*, 1010–1027.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*, *12*, 742–756.

U.S. Department of Health and Human Services. (2015). *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. Retrieved from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, *3*, Article 1. doi:10.1525/collabra.74

Vinkers, C. W., Finkenauer, C., & Hawk, S. T. (2010). Why do close partners snoop? Predictors of intrusive behavior in newlywed couples. *Personal Relationships*, *18*, 110–124.

Vision, T. J. (2010). Open data and the social contract of scientific publishing. *BioScience*, *60*, 330–331.

Wicherts, J. M. (2016). Data reanalysis and open data. In J. Plucker & M. Makel (Eds.), *Toward a more perfect psychology: Improving trust, accuracy, and transparency in research* (pp. 215–232). Washington, DC: American Psychological Association.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728.

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology*, *12*, 313–325. doi:10.1007/s10676-010-9227-5