Running head: SHARING NON-INDEPENDENT DATA

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

Samantha Joel
University of Utah

Paul W. Eastwick
University of California, Davis

Eli J. Finkel
Northwestern University

In press, *Advances in Methods and Practices in Psychological Science*

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 recent speed-dating article we published in *Psychological Science* (Joel, Eastwick, & Finkel, in press), along with a summary of the issues we considered and addressed along the way. As we drafted the present paper, however, the third author felt unsure in retrospect about some of the procedures we followed, especially if our approach were to be perceived as a model for open-data decisions in other, more typical cases involving non-independent data. This paper addresses these concerns, but also identifies areas of consensus. All 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 non-independent 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.

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

Psychological scientists face a pressing need to improve the transparency of our research. Within the last decade, evidence has been building that the reproducibility of psychological findings has considerable room for improvement (e.g., Open Science Collaboration, 2015; Munafo, Nosek, Bishop, Button, Chambers, Percie du Sert et al., 2017; 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, Conner, De Fruyt, De Houwer, Denissen, Fiedler 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 results do 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 re-analysis and meta-analytic aggregation over time (e.g., Wicherts, 2016); in many cases, scholars can use the datasets to conduct novel analyses that the original researcher(s)—who collected the data—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 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. With social psychology, 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 close others in their lives, and willingness to harm others, among other topics (see Gilovich, Keltner, Chen, & Nisbett, 2016). When participating in research studies such as these, participants place a great deal of trust in the researchers to protect their confidentiality; indeed, without guaranteeing confidentiality, 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 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 their willingness to share their personal experiences with us.

Protecting confidentiality is not always as straightforward as it might seem. In theory, one can de-identify a dataset by simply removing personal information (e.g., names, email addresses) before making the data publicly available. However, variables that do not appear to be personally identifiable can sometimes be used, especially in combination with other data, to re-identify a dataset (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 collected. 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 identifies 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 non-independent data: data for which observations are nested within groups, such as couples or families. For an independent dataset, one must ensure only that the participants' responses cannot be identified by an outside observer. However, it is uniquely challenging to de-identify non-independent datasets in a way that protects participants' information *from each other* (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 a participant cannot discover via a "confidential" study that their partner does not love them, or is no longer attracted to them, or still pines for an ex, etc.

Overall, the open sharing of data offers a powerful way to increase the reproducibility and replicability of our findings, as well as the overall contribution of the data to the scientific community. Yet, if those of us in fields that rely on data that cover sensitive topics (e.g., relationship science) wish to reap these benefits, we must develop and adhere to procedures for sharing our data that also protect participant 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 non-independent 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 robust literatures already exist in that space). We offer a detailed discussion of how we were able to circumvent these complications in a recent case that concluded with a dataset made openly available to researchers.

Our Own Foray into Sharing Non-Independent Data

We—the authors of the present report—recently underwent our first attempt at openly sharing sensitive, non-independent data. Specifically, we sought to share the two speed-dating datasets required to reproduce the findings in Joel, Eastwick, and Finkel (in press). In these studies, 350 participants completed, in 2005 or 2007, a long intake questionnaire relevant to mate selection (i.e., traits, ideal partner preferences, and other individual differences measures). Participants then attended a heterosexual speed-dating event, which included approximately 12 men and 12 women. At the event, each participant had 4-minute speed-dates with each opposite-sex participant (total $N = 2050$ speed-dates), followed by a questionnaire on which they reported on that speed-date. 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 others to reproduce our findings, we needed to share data from over 100 self-report measures collected from each sample.

As elaborated below, we consulted various sources and considered several options before choosing the data repository UK Dataservice (www.ukdataservice.ac.uk). Data uploaded to this repository are stored in a timestamped, noneditable, and nonretractable format. We used the UK Dataservice's "safeguarded access" option, which requires researchers to register an account on the website (free of charge) before accessing the data. Registration requires a stated affiliation with a suitable professional organization (e.g., a

university) and a valid, accompanying institutional email 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 non-independent 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 above) 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; as such, this part of the paper 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 others should follow vis-à-vis their own datasets. We describe the process that we used before turning to Eli's retrospective reservations about it. Paul and Sam offer rejoinders to Eli's concerns, and we conclude with a consensus section oriented toward maximizing the public sharing of data while minimizing risks to participants.

The Process We Used

This section reports on the three primary questions we asked ourselves when evaluating whether it was acceptable to post our data. It also provides the answers we developed in response to these questions.

1. Is anonymization possible?

We first made every effort to de-identify the data. As these were non-independent data, we carefully considered the identifier variables that link 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 these variables, people could conceivably use their knowledge of their own responses to identify others' responses. In our studies, for example, a participant 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 attractive, desperate, etc. specific other participants in the study perceived them to be.

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 individual item responses wherever possible. Specifically, we removed any variables that indicated membership of 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 individuals of a certain age, ethnicity, etc. to participate). We also took the unusual step of removing identifier variables from the dyadic dataset: Although identifier variables are typically required to reproduce analyses that account for non-independence (e.g., in multilevel models), our particular analytic strategy (machine learning) dealt with non-independence prior to data analysis. That is, our analyses accounted for non-independence through the centering of the dependent measure rather than with the use of identifiers, and so the identifiers are not required to reproduce the analyses reported in the manuscript. Our final materials still included gender and a large number of individual item responses, 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 importantly, (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 10-12 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 non-independent dataset that contains identifier variables, the data must be anonymized to the point that the likelihood of a participant identifying their *own* responses would be miniscule. This could potentially be accomplished by removing not only demographic information, but also all individual and categorical items for which a participant might conceivably recall how they rated the items and thus recognize their own data. Ideally, the dataset would be left with only centered and/or aggregate variables that have no extreme outlier responses, such that it would be impossible for a participant to discern which responses belonged to them (and, by extension, which responses belonged to their partner, friend, etc.). Alternatively, the dataset could be shared in a repository that effectively prevents participants from accessing the raw values from the dataset. We return to a discussion of repository options later in the paper.

2. Did the consent process address data sharing?

Participants may have a variety of preferences surrounding the sharing of their data, both in favor (e.g., to maximize the contribution of their time and effort) and in opposition (e.g., concerns about confidentiality and sensitivity). Ideally, these preferences should be considered during the consent process (Cummings, Zagrodney, & Day, 2015). As our current paper relied on existing datasets—with consent taking place in 2005 or 2007—we verified with the Northwestern IRB that sharing these datasets on the UK Dataservice 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 deidentified, 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 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 we had included this language in the consent form 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 Dataservice 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", ICPSR, 2017).

3. How great is the potential for harm?

Our third consideration was the *degree* of sensitivity of the data. What are the risks associated with linking people's identities to their responses to the specific measures we collected? Our assessment was that these risks were reasonably low in our case; although our studies included a number of somewhat delicate or embarrassing measures (e.g., desperation to find a romantic partner, others' perceptions of own 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 10-12 years old at

the time we sought to share them, even if anonymity were breached (which we deemed extremely unlikely, see above), we viewed it as unlikely that any openly shared data would be harmful enough to meaningfully affect anyone's ongoing relationships. After carefully considering and discussing these issues, we concluded that the "safeguarded access" option from the UK Dataservice sufficiently mitigated any remaining risk.

Retrospective Wariness: Eli's Squeamishness about Relying on the Subjective Judgments of Individual Researchers to Make Ethical, Technological, or Legal Judgments about When Sharing Data is Appropriate

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) 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.[1] 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 of 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 less confident than I was in 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

[1] This section is written from a first-person perspective to bolster narrative efficiency.

relevant issues are complex in every case, but much more so in the vast majority of cases involving non-independent data than they were in our case (because, as noted previously, our machine learning procedures allowed us to exclude a group-level identifier variable). As such, I am not confident that our approach should serve as a model for others. My hope is that going public with the internal debates that Sam, Paul, and I had will be useful for other scholars grappling with these sorts of issues.

Regarding *anonymization*, we had concluded that the likelihood that participants would identify their own responses in our dataset is small, and that the likelihood that participants will identify other participants' responses is minuscule. But data hacking and computer security are major industries, and a clear conclusion from experts in that space is that apparently secure data often are not secure (Zimmer, 2010). For example, if a participant recalls reporting, on the speed-dating intake questionnaire, that she said 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 was hers. As I reflected more deeply on these issues, I began to wonder: By what criteria are psychological scientists (ourselves or other people) 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 Dataservice safeguarded access option." But, upon reflection, the meaning of "the academic community" is ambiguous. What it actually means for the UK Dataservice is: anybody with an institutional (e.g., .edu) email 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 email

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 in many studies for participants to access their partner's raw data.

Regarding *potential for harm*, we had concluded that 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, others' perceptions of own 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)." 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 feel these days to be sensitive. What criteria should a psychological scientist use to determine whether data are too sensitive to share, especially in situations where 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 from the *Psychological Science* paper 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 dataset had some extremely rare features—especially that the data were more than a decade old and that we were able to 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 the procedures we used—answering self-interrogations based on our best intuition—may be excessively risky for the vast majority of non-independent datasets 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—"answering self-interrogations based on our best intuition"—in a slightly broader context.

For example, here are two strategies I like less: (a) Mindlessly and automatically post your nonindependent dataset for the general public to access, or (b) never even consider posting your nonindependent dataset 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 nonindependent data sharing, worked hard to address those risks, and posted the data at the end of this long process.

Here is a strategy I like more: Ask a group of data science professionals to evaluate whether we had properly anonymized our dataset. 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 more training in data anonymization than the (limited) training that I and my coauthors have received (see also Meyer, in press). Overall, I found it frustrating and disappointing that so much digital ink has been spilled over the importance of data sharing, and very little of it has been devoted to helping researchers with (modestly) complex datasets 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 beats both the mindlessly-share and obstinately-refuse-to-share strategies.

Rejoinders (from Sam)

Eli raises many valuable critiques that the field must consider as we move forward with open practices. I would like to offer two rejoinders concerning the risks associated with generalizing our procedure to other cases. First: while the data for our machine learning paper 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 a variety of raw, individual items. In more normative cases, reproducing analyses frequently requires only aggregated and/or centered variables. It would be considerably more difficult—if not impossible—for participants to identify themselves from a dataset containing no raw values. If it is not possible for the participant to use their knowledge of their own responses to identify themselves, then they cannot use the identifier variable to identify others, even if that identifier variable remains in the dataset.

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. For researchers looking to share their data, the question they must answer is the extent to which a given *combination* of safeguards protects confidentiality in the context of their particular study. For example, in the case of a non-independent dataset that has been de-identified and shared in a vetted repository, one must consider the odds that a given participant has a unique outlier response that they remember providing, *and* is able to use it to identify a confidential response from another participant, *and* is able to meet the repository's vetting criteria by the time the data are shared, *and* is motivated and resourceful enough to sleuth out 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 participant 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 to 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.

1. **More, and better, data repository options**

When researchers cannot make their data fully open due to ethical considerations, the journal *Psychological Science* recommends using "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 (and only qualified professionals) who requests access. As of April 2017, we found a dearth of data repositories that offered this service. The SAGEPub-recommended search site http://www.re3data.org/ yielded few options, as the vast majority of databases allow a researcher to either make their

data fully public or keep their data fully private, without 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 Dataservice, which is the repository that we ultimately used. Two other promising suggestions we received were OSF.io and the Harvard Dataverse; however, neither currently offers any level of third party vetting. OSF.io also does not yet offer a way to register, or time-stamp, data, yet also keep it private indefinitely; embargos are automatically released after a maximum of four years. Another suggestion we received was ICPSR (https://www.icpsr.umich.edu/). This website charges the requester $350 and also requires the requester to obtain IRB approval before granting access to the data. The data are then mailed to the researcher on a compact disc. Given these barriers to access, this repository seemed too restrictive for our purposes. However, this website may offer a useful compromise for highly sensitive studies (see above).

The first author of the present paper also inquired about the issue of non-independent data at a seminar on scientific integrity (Nelson, Simonsohn, & Simmons, 2017, April). The presenters had the innovative suggestion that an application could be created, perhaps using Shiny Apps (https://www.shinyapps.io/), that would allow researchers to run analyses on a particular dataset 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, a "mimicked" dataset could be generated that reproduces the central features of the real dataset (e.g., using the R package synthpop) and be made 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 Dataservice was the only appropriate (albeit imperfect) repository we found for our current 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, American Psychological Association) 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 the dataset, considerable thought needs to go into the vetting criteria.

Data sensitivity guidelines and training

At what point is a dataset sufficiently deidentified 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 when making the decisions described above, and it is possible that our intuitions could have been wrong. Data science is a vast field of inquiry, and if we want psychology to be a mature, 21st century discipline with respect to technological sophistication and transparency, we must tap into that expertise.

One route to achieving this would be for us to develop our own training. Currently, it is rare for psychology graduate programs and statistics classes to include training on data

anonymization, despite 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 deidentify a dataset, akin to those that are provided to health researchers (e.g., Health and Human Services, 2015), particularly in fields within psychology that frequently work with sensitive data.

Another option, especially in the near-term and in the absence of our own established protocols, would be for us to outsource the task of ensuring confidentiality to trained professionals. Universities and/or professional organizations like APS or APA could establish a service whereby data scientists are available to double-check datasets with respect to sensitivity and anonymity before the research team shares them.

2. Protocols to increase compliance with data sharing requirements

Despite a researcher's best efforts, there may still be datasets 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 (American Psychological Association, 2002). However, this requirement is unenforced and adherence to it is poor; one team of researchers was able to obtain the original data to 64 out of 249 datasets published in 2004, with 73% of authors failing to share their data (Wicherts, Borsboom, & 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, with 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 wherever possible because it allows researchers to access data anonymously, eliminating status and reputation concerns for the requester.

Conclusions

Non-independent 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 that are able to vet or protect access to data, potentially hosted by universities or professional associations, (b) accessible resources 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.

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., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & 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 [Blog 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,* 1-11.

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* (Fourth Edition). New York, NY: Norton.

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

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. Message posted to facebook.com/groups/psychmap.

Joel, S., Eastwick, P. W., & Finkel, E. J. (in press). Is romantic desire predictable? Machine learning applied to initial romantic attraction. *Psychological Science.*

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

Meyer, M. (in press). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science.*

Munafo, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., Loannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature: Human Behaviour, 1,* 1-9.

Nelson, L., Simonsohn , U., & Simmons, J. (2017,April). *Scientific integrity.* 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.

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, C. J. P., Johnson, S. L., Krueger, R. F., Miller, J. D., Oltmanns, T. F., & 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.

Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology, 3.*

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

Wicherts, J. M. (2016). Data re-analysis and open data. In J. Plucker & M. Makel (Eds.), *Doing good social science; Trust, accuracy, and transparency.* Washington: American Psychological Association.

Wicherts J, 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. http://dx.doi.org/10.1007/s10676-010-9227-5