AERA Open
January-December 2021, Vol. 7, No. 1, pp. 1–11
DOI: 10.1177/23328584211006475
Article reuse guidelines: sagepub.com/journals-permissions

Data Sharing in Education Science

Jessica A. R. Logan
The Ohio State University

Sara A. Hart
Christopher Schatschneider
Florida State University

Many research agencies are now requiring that data collected as part of funded projects be shared. However, the practice of data sharing in education sciences has lagged these funder requirements. We assert that this is likely because researchers generally have not been made aware of these requirements and of the benefits of data sharing. Furthermore, data sharing is usually not a part of formal training, so many researchers may be unaware of how to properly share their data. Finally, the research culture in education science is often filled with concerns regarding the sharing of data. In this article, we address each of these areas, discussing the wide range of benefits of data sharing, the many ways by which data can be shared; provide a step by step guide to start sharing data; and respond to common concerns.

Keywords: *data sharing, open science, data management*

Researchers in the education sciences may have noticed that an increasing number of researchers, journals, and funding agencies are discussing the idea of open science practices, including the idea of sharing data. Data sharing has entered the conversation in the United States in part due to the 2013 requirement that any data collected using federal funding are mandated to be open and accessible to the public (The White House, 2013). In response to this federal order, four major federal agencies that fund education sciences research (National Science Foundation, National Institutes of Health [NIH], National Institute of Justice, and the Institute of Education Sciences [IES]) now require applicants to include explicit data sharing plans as part of their grant applications. For IES-funded projects alone, there are approximately 350 research projects that have been funded that are subject to this data sharing requirement, all of which should have data that are ready to share (Albro, 2020).

Paradoxically, the practice of data sharing in education science remains relatively rare. A recent survey found that education researchers believe data sharing to be the least common of the open science practices in education (rating it as less common than open access, preregistration, open materials, or replication), and report never or rarely engaging in data sharing practices themselves (Makel et al., 2019). Many researchers are not trained in the act of sharing their data, are uncertain how to incorporate such practices into their work, and are hesitant to participate in the practice due to some nagging concerns that are often vague. Most are unaware of the benefits of sharing data both for themselves and for the larger enterprise of science. Therefore, the purpose of this article is to describe the importance and practice of data sharing in education research.

Though the importance of data sharing for education research has been described in previous work about global open science practices (Cook et al., 2018; McBee et al., 2018), this article is unique in its depth of coverage on data sharing and is specifically written for education researchers who are new to the practice (see also Meyer, 2018, for an excellent review of practical tips for data sharing in psychological science, many of which apply here). In this article, we will define data sharing, highlight why the practice of data sharing is important, walk through the process of sharing data, and list and address some common data sharing concerns held by education researchers. The authors all primarily follow quantitative lines of inquiry, and therefore, our discussion is focused on quantitative data. (However, see Tsai et al., 2016, and Mannheimer et al., 2019, for general information about qualitative data sharing.)

What Is Data Sharing?

Data sharing refers to the process of taking any type of research data and making it available for other researchers to examine or use. Most researchers might not realize that they

have already been participating in data sharing in their published papers. It is commonplace to share processed data, such as means, standard deviations, sample size, correlation matrices, scatterplots, or tables of results and coefficients. This type of processed data can be helpful for meta-analysts and other niche data uses (e.g., using a correlation matrix as the input data for a new structural equation modeling analysis), but in most instances, data sharing is referring to the sharing of participant- or variable-level data, not simply summary statistics.

Readers may be familiar with the "data are available upon request" disclaimer, where corresponding authors agree to provide access to the analytic dataset used in a particular paper. This type of data sharing does technically meet most federal guidelines (e.g., NIH, IES) and may seem to be a simple solution to data sharing requirements. However, sharing data on request also has several disadvantages, both for the original data holder and for the potential requester. When presented with a request, the data holder must spend time and resources preparing the data for sharing, answering the requester's questions, and learning and complying with the institution's most current rules and regulations around sharing data (this often includes executing a formal data sharing agreement). This may be why, historically, many authors have been unresponsive to such requests (Wicherts et al., 2006). From the perspective of the requester, research studies have suggested that requesting data access is often prohibitively time consuming (e.g., Langille et al., 2018), and the availability of the data declines over time (Vines et al., 2014).

Each of the mentioned disadvantages can be mitigated by planning, proactively, to share collected data formally through use of a data repository. Using a data repository can democratize access to research data beyond internal research networks. In addition, data repositories typically guarantee preservation of the data over time, and allow access to the data, and data documentation, beyond most investigators' memories and computer lifespans. Data repositories can also be used to store the data not associated with one specific paper, but as a final dissemination product of the entire research project. We note that this is a final product, because even if data sharing is required by your funding agency, they do not expect that the data will be released as quickly as it is collected. Rather, it is more common for datasets to be shared after one or all the primary aims have been published. Because this sharing of data in a data repository (vs. "data available upon request") is less familiar to education researchers, and is preferred by some newer federal guidelines (e.g., NIH, 2020; effective 2023), the remainder of this article will focus on the use of an established data repository to preserve and share data.

Successful Examples

Open data enables new intellectual possibilities for data already collected, exponentially increasing the return on the time, effort, and money already invested into any given dataset. For example, the field of astronomy was an early adopter of data sharing and openly archived the data from the Hubble Space Telescope. Of the peer-reviewed publications using the data from the Hubble Space Telescope, 60% (500) were from investigators who were not the primary investigators or their collaborators (Ember et al., 2013). In addition to the over 200 peer-reviewed publications by the primary investigators and their collaborators from these data, there was still enough intellectual capital in the archived data to enable the publication of more than two times that number of the primary papers.

An example from developmental science is the Child Language Data Exchange System (CHILDES). CHILDES is an open data repository language acquisition corpora data, established in 1984. CHILDES started with a small group of researchers agreeing to share their corpora data. As of last count, there are now over 5,000 researchers who have joined CHILDES as members, the data repository has over 110 million words published, and over 7,000 papers have been published citing data drawn from CHILDES (MacWhinney, 2000, updated in 2020 under the same citation). Although there is no way to know what the original investigators of these corpora would have published without contributing to CHILDES, the impact on the scientific field those investigators have had by sharing their data in CHILDES, based on the CHILDES citation numbers, is remarkable. It is estimated that at any given time, 100 research labs around the world are working on a project using the CHILDES repository (MacWhinney, 2000, version updated in 2020).

These two examples of community-led large-scale data repositories[1] are a best-case scenario; they represent an extreme high end of data reuse. In reality, there are very few publicly shared datasets that will have similar rates of sharing. However, given that data sharing, and the public mandates to support data sharing, are still relatively new, we believe that data sharing and data reuse are still in their initial growth phase. In the next section, we will describe some of the advantages of data sharing, including why granting agencies are so keen for you to start.

Why You Should Want to Share Your Data

Why have granting agencies decided that data sharing is necessary for all federally funded projects? Their reasons are nicely summarized by the new NIH Policy for Data Management and Sharing: "Sharing scientific data accelerates biomedical research discovery, in part, by enabling validation of research results, providing accessibility to high-value datasets, and promoting data reuse for future research studies." This quote lays out a framework to discuss the many benefits of sharing your data. The statement "enabling validation of research results" points out that data sharing allows for other researchers to use the data that you

share in order to check your published work. Thanks to the credibility revolution, or replicability crisis (Vazire, 2018), the idea of greater transparency in the research and publication process has been a major driver in the shift toward data sharing. By sharing your data, as well as data documentation and analysis scripts, you allow others to replicate or extend analyses that you have already published. This can take the form of direct replication, as robustness checking through the addition of new covariates to your published models, or as scrutiny when new methods or analytic techniques are used to reexamine or pose questions that have already been asked of the dataset. As established in 2018, different analytic choices can show different results, even within the same dataset and using the same variables (Silberzahn et al., 2018). Such examinations can test whether the findings you report in one paper are robust in multiple samples, with multiple methods, and in multiple contexts, strengthening the theoretical contribution to the scientific space.

Data collected by specific researchers or research teams always have the potential to be biased toward their proposed hypotheses (Wicherts et al., 2016). When working with several external datasets, there is less of a chance of such researcher degrees of freedom influencing the results. Conducting your analysis twice, once on the data you collect and once on the data you have pulled from a repository, can provide a stronger argument for the potential findings and widen the contribution of the published work. As an example, in a study of children's television viewing behavior, Khan et al. (2017) presented an analysis of their actively collected research sample and then also included an analysis of the same question in a nationally representative dataset, the Early Childhood Longitudinal Study Cohort. The ability to check whether and to what extent a hypothesized result holds in a second dataset strengthens the argument for the external validity of new research findings. Through comparing datasets, researchers can test whether an effect is consistent across multiple studies that tested similar ideas, thereby strengthening the research base.

In the NIH statement, "Providing accessibility to high-value datasets" points to the idea that data sharing allows for a more democratic sharing of research resources toward the goal of supporting the scientific process. This is the principle of beneficence, which was established in the Belmont report (U.S. Department of Health & Human Services, 1974), and states that scientific work should always maximize possible benefits of any research conducted. By sharing the data you have collected with other scientists, you are maximizing the possible benefits by allowing the data to be reused.

In another example of how data sharing can result in equitable accessibility of high-value datasets to the benefit of science, data sharing can allow a researcher to combine across datasets to answer research questions that are not possible with a single dataset due to low numbers of participants or low occurrences of a given behavior of interest (Curran & Hussong, 2009). For researchers who work on rare disabilities, or do not have large data collection budgets, it might not be feasible to collect enough data to conduct advanced statistical modeling. By combining datasets, a researcher can potentially achieve the statistical power needed for modeling, while spreading the research investment across multiple researchers. Finally, even if the datasets cannot be combined due to lack of measurement overlap, effect sizes of interest could be generated for each study and combined via a meta-analytic framework.

The NIH further states "promoting data reuse for future research studies" because the data you collect in the service of a particular research aim can be used to answer questions that you may not have ever considered. Data sharing facilitates research via data reuse, which includes testing new ideas that the original investigator may never have considered; data sharing not only makes science more economical but also advances the pace of scientific understanding (Freese, 2007; Munafò et al., 2017; Vision, 2010). Opening data to others provides the mechanism and opportunity to allow many other people to test out their new ideas.

This brings up another excellent reason for data sharing: It opens the resources you have to others. This is particularly helpful to researchers who are at early stages of their careers, or who are working at institutions or organizations that do not have the infrastructure to support large research projects. It similarly supports those researchers without a network of well-resourced scholars, a situation that disproportionately affects underrepresented researchers (scholars who are Black, Indigenous, or people of color) in the academy. Data sharing can be part of your antiracist work.

Finally, sharing your data can also benefit you as the primary investigator directly in other ways. Research suggests that papers published with open access to data have an increased citation rate (Drachen et al., 2016; Piwowar et al., 2007; Piwowar & Vision, 2013). A recent analysis of publications with data provided in a repository (without restriction) had up to a 25% higher citation rate (Colavizza et al., 2020). In addition, published datasets themselves are also citable. Datasets are published with a digital object identifier (DOI), which will be cited by anyone who uses your data in a subsequent paper or analysis. You can use these citations to support the effectiveness and reach of your work on your grant reports, applications for funding, annual reviews, or promotion and tenure packages. It can be exciting to see your research products be used in ways that you have never anticipated.

Principles of Data Sharing

There are four internationally accepted features of good data sharing, called the FAIR data principles (Findable, Accessible, Interoperable, and Reusable; Wilkinson et al., 2016). Most of the key data sharing concepts can be classified into one of these four categories. In this next section,

we will discuss each in turn and what they mean for a data sharer. We will then follow with a detailed guide on how to share your data.

Findable

To be able to reuse data, someone needs to be able to find them, the findable data principle. Therefore, when sharing data, data needs to be accompanied by rich metadata in order to assist the findability of the data. Metadata are the data that describe the data. Metadata allow a user to know what is inside the dataset without opening it. Think about this as you would keywords for an article or search terms in a library. For your article, your keywords serve the purpose of helping other scholars find your work and know the key topics and points. In education and developmental science, good metadata would include information about the study design (e.g., experimental or correlational), the participants (e.g., age, sample size, population drawn from), and the variables (e.g., general construct information, number of time points).

Accessible

Once a user finds data, they need to be able to access it. Minimally, anyone with a computer and internet should be able to access at least descriptive information about the study and dataset (the metadata). Therefore, as the data sharer, we suggest that you store your data in a data repository. A data repository is any place where you can store your data and accompanying metadata, and that provides access to others. The accessible principle does not necessarily mean that all data need to be "open" (i.e., freely available to anyone from the internet), although we encourage that. Instead, data are accessible if the metadata are openly available, and if a user knows how they can apply or otherwise request access to the actual data.

Interoperable

The third principle of reusing data is making it interoperable. This is the idea that data and corresponding metadata need to be stored in a way that other computers can read it. If you have ever tried to share a data file with a collaborator who uses a different statistical software package than you do (e.g., you use SPSS and they use Stata), you have experienced firsthand the issues of interoperability. Sometimes, files saved in proprietary software formats can even become unreadable between versions (i.e., lack of backwards compatibility). It is for this reason that we recommend that you store your data and metadata using formats that are not specific to a certain proprietary software (e.g., .sav files from SPSS, .pdf files from Adobe), but instead are general formats readable by all software (e.g., .ASCII, CSV, tab-delimited files, and .txt for document files). Any statistical software will have the capability of saving or exporting your dataset as at least one of these file types.[2]

Reusable

The reusability of data contains two important concepts. First, anyone who does find and access your data also needs to have all the critical information that will help them understand your data. This again comes down to high-quality metadata. Good metadata not only describe the data it accompanies (like discussed in the Findable section) but also describe the broader project, giving context to the data. We will describe this in more detail in the next section. The second concept is data and metadata provenance. Make sure that all data and data documentation have DOIs, with authors and other important information clearly assigned, allowing the data and accompanying products to be properly cited. You should also assign an Open Data Commons (https://opendatacommons.org/) or Creative Commons (https://creativecommons.org/) license, which will tell data users exactly how they can use and cite your data and data products. The generation of these provenance measures is typically a built-in part of the data depositing process at any online data repository.

How to Share Data: Step-by-Step

For those projects that you are preparing to share data, here is a brief step-by-step guide to follow. Investigators can follow these steps in order, starting at the beginning of their project planning or at any point the data have already been collected, and successfully share their data.

Before the Study Starts: Informed Consent Check

When it comes to data sharing, we will first consider your informed consent. There are at least two times in the research cycle you should consider your informed consent. For investigators starting new projects with the hopes of sharing data in the future, consideration of eventual data sharing should be done during the planning stages before data collection occurs. We highly recommend you reconsider all the informed consent language template you have previously used (i.e., do not copy your informed consent text from a previous project[3]). In our experience, investigators wrongly believe that they know the language institutional review boards (IRB) "want" to see concerning data storage and sharing, and often, their beliefs are much more restrictive than what is currently true (e.g., thinking you need to promise to destroy your data, or promise to not share, or promise to analyze only a small set of research questions; Meyer, 2018). We recommend Meyer (2018) for a general review of best practices in data sharing, Shero and Hart (2020c) to assist investigators when approaching their IRB, Shero and Hart (2020b) for an IRB protocol template,

Shero and Hart (2020a) for an informed consent template, and a Center for Open Science collection of templates and resources (https://osf.io/g4jfv/). We also note that IRBs are given the latitude to make their own rules as long as they meet current guidelines and standards (e.g., the International Compilation of Human Subjects Standards, The Common Rule), and therefore, it is impossible to provide general rules of thumb about IRBs. Therefore, we recommend you talk with your IRB and get advice about their data sharing policies.

If you have already collected your data, then before you share any data, you should review the informed consent documents that your participants agreed to. This can be done at any point after the data collection is completed, including years later.[4] Many research labs use templates or historical documents used by previous labs, which can include language that is restrictive to data sharing. As you review, be on the lookout for phrases such as "These data will only be shared with study investigators" (this could possibly restrict your ability to openly share your data), or "All data will be destroyed after 7 years" (although surprising, this is not uncommon language, and could mean something you did not intend). If the informed consent does have language you are concerned about regarding the possibility of data sharing, do not despair. You should start a conversation with your IRB to determine what your informed consent will allow you to do regarding data sharing, and if they would consider allowing a waiver of consent to allow data sharing.

Before the Study Starts: Data Entry

A critical part of the research process is determining how data will be captured and digitally organized. Data entry is a complex topic that is important for data accuracy, and a complete discussion is outside the scope of this work (see Burchinal & Neebe, 2006, Logan, 2019, and Reynolds & Schatschneider, 2020, for some detailed guidelines). However, two key data entry specifications are important to mention here.

Item-Level Data. Many education science projects collect survey or test data that need to be scored, added together, or otherwise processed prior to analysis. For example, a measure of the school climate may be calculated as the mean score across 12 manifest items completed by a teacher, or a third-grade vocabulary test scored as the number of items the student answered correctly. With data like these, we recommend that participant responses to each item be captured and documented in the database (i.e., item-level data). This maximizes the possibilities for data reuse, as item-level data can be combined in different ways following different research questions, can be subjected to different measurement models, and can be used in integrative data analyses (Curran & Hussong, 2009).

Variable Names. When designing your database, create a general rule for how variables will be named. Variable name conventions can be developed to include information about several facets of the collected data. These include the tool used to gather the data, when the data were collected, who the informant was (e.g., which parent reported), about whom the data was collected (e.g., the lead teacher or coteacher), and/or whether the variable is an item or a total or summary score (Burchinal & Neebe, 2006). Variables should be named consistently, and uniquely, across all datasets associated with a project (Reynolds & Schatschneider, 2020). When variables are named systematically, it improves the transparency of the data and so minimizes barriers to entry for data reuse.

After Data Have Been Collected: Clean Your Data

After data collection, but before data can be shared, you should check them for errors and clean them. This is an important step for you as the investigator to undertake, as you know the parameters of your data in ways that a new user never could. Here we present two critical data cleaning steps to consider.

First, check the rate of missingness on each of your variables to make sure that it is as expected. You might expect missingness when it is planned in the design (e.g., skip logic) or if some items cannot be given to some participants (e.g., children under age 4 years do not receive the vocabulary assessment). You may also have known events that cause data loss—for example, a school building may have closed such that you were unable to collect their data. More commonly, participating teachers, families, or children will leave the study over time. Comparing the missingness rates observed in the data with those documented through other tracking methods will alert you to any unexpected missingness (e.g., the stack of assessments a research assistant accidentally filed without entering the data).

Second, conduct range checks to ensure all of the observed data points are within the expected and possible range of values. Make sure to check this for all your variables, including individual items, total scores, and dates. For example, in data you recently collected on kindergarten children, no one should have a birthdate from 1979. Similarly, if a response variable only has two possible values (e.g., correct = 1, incorrect = 0), then no cases should have a value of 3. To conduct these range checks, examine the minimum and maximum values for each variable in the data.

After Data Have Been Collected: De-identify Your Data

After data collection, but before data can be shared, an investigator needs to consider data de-identification. With any data on human participants, we must always be concerned with protecting the confidentiality of our participants. We can do this through data de-identification (for other

resources, see https://venngage.net/ps/5p6yjaAGTSs/new-5-things-to-check-for-data-de-identification and Edwards & Schatschneider, 2020). Investigators should identify and remove Personally Identifying Information, defined by Health Insurance Portability and Accountability Act (https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#standard), which includes identifiers such as names, birthdates, social security numbers, and location information, and limit the number and detailed breakdown of categories within variables on the file (see Federal Committee on Statistical Methodology, 2005; U.S. Department of Health & Human Services, 2012). When studying children, age is usually a particularly important variable, so consider converting birthdates and test dates to a calculation of age at each collection point. For many studies, this de-identification step will be enough. However, if your data are at heightened risk of being sorted or disaggregated in such a way that a single person could become identifiable, there are several additional strategies you can employ to increase confidentiality.

Each of the following strategies can be used to further reduce the likelihood that a single case will be identified in your dataset (see U.S. Department of Health & Human Services, 2012, guidelines for additional details). First, you can use truncation (top or bottom-coding) to recode or truncate any extreme values so that no single person has the lowest or highest value, but instead, perhaps, the lowest or highest five people in the study receive the same score. As a similar option, you can use rounding to recode values into larger bins across the entire distribution, so no case is the only case with any given value. Another method is to deliberately introduce noise—adding or multiplying the observed scores by random numbers so that each subject's exact score cannot be known. A similar noise-introducing strategy is to randomly replace one individual's reported value with the average of their small group (also called blurring). Finally, rank swapping (also called switching) is particularly useful when geographical information (e.g., school or district) might be ascertainable. In rank swapping, subjects across schools might be matched on all relevant variables and then swapped, with the idea that some of the subjects in a school may not in fact be from the school, but their data are close enough to represent the student being swapped.

The overall goal of these strategies is to try to ensure confidentiality. If you still feel that the risk of sharing your data is too great, or if there is no way to conceivably protect confidentiality, one possibility is to not share any individualized data. Instead, you can choose to only share summary statistics such as means, standard deviations, sample sizes, and covariance matrices, and include them broken down by as many subgroups as possible. Even these summary statistics can provide opportunities for reexamination and reanalysis.

Preparing Data for Sharing: Document Your Study

Data can only be reused if the person reusing them understands the purpose and methods of your original study. Others will not know the strengths and limitations of the data without knowing the context in which the data were collected. Good reusable data require good metadata. Although this step can be done after data collection is completed and just before data sharing, we recommend that investigators set up their process to start their study documentation at the very start of their project planning. It is very easy to forget the details. However, it is likely that the completion and finalization of the study documentation will happen just before data sharing. Study documentation metadata, also called codebooks, are composed of several important pieces.

Summary Documents. Write a brief overview of the study background, aims, and data collection process. Think about this like the one-page summary document that you might compose when you submit a grant.

Information About the Sample. In this section, include all information relative to the recruitment, ascertainment, and retention of your participants, and the timeline of your study. Relay how the sample was selected. Document any exclusionary or inclusionary criteria, as well as why those criteria were selected and how many potential participants were eliminated for not meeting them. Include a CONSORT diagram, which provides a framework for documenting how participants are recruited and maintained in a study (see www.Consort-Statement.org). Though designed for longitudinal interventions, the CONSORT diagram is an excellent tool for summarizing sample information.

Study Protocols. The study protocols are in-depth descriptions of all study procedures. This will be the majority of the submission. In education science, this will probably include how testers or observers were trained, detailed information about the measures used, and the data collection processes. Some data may be collected by trained assessors, others through observation or parent/teacher report, and this should be documented in this section. If not completely raw data, describe how the data were intermediate processed (e.g., sum scores were calculated by combining which variables and how). Additionally, describe how the data were entered, how they were scored (when applicable), any changes you made to the raw data (e.g., eliminating observations without valid scores), and whether and to what extent the data were checked for accuracy.

Measures. Descriptions of the measures are also needed so that other users will know exactly what specific assessments were included. This also includes "project made" assessments like demographic questionnaires and other tools that

were designed by your team. For each measure given, you should document the name and publication information (either the publisher or a relevant study where the measure has been used before). More details about the assessments given would also be useful to others, such as the minimum/maximum possible scores on assessments and any decision rules that your project employed to score these assessments. Inclusion of when the data were collected is also critical, both in terms of time of year, and at what stage of the study (e.g., screening, pretest, posttest).

As part of this section, also document any changes that were made during data collection. For example, imagine a survey was changed halfway through collection to add 12 new items. That will mean that those 12 items will have missing data for some subset of the sample. Let the potential data users know that the change was made and that this amount of missing data is expected. To consider how to document missingness more generally, see the next point.

Document Missingness. Make sure that any missing data are clearly defined and documented. Different projects accomplish this in different ways. Some projects use discrete values (e.g., −99), or insert a value such as N/A, while others leave missing cells blank. Be aware that missingness may not be coded in the same way across all variables in your dataset. For example, if a respondent skipped an item on a survey, you might have left the missingness as blank, but inserted a code of −999 when the response is missing due to skip logic. While it is tempting to code every possible reason that data might be missing as part of the primary variable (e.g., −999 when data are missing due to a skip logic, −998 when the data are missing because the participant was too young to receive an assessment, −997 for a child who moved out of town), such codes are rarely informative during the analysis phase.

Data Dictionary. The data dictionary is a shorthand overview of each variable, organized by variable name. Data dictionaries provide a one-to-one linkage between the variable name and what that variable stands for in the data. If the variable stands for assessment data, it should describe what kind of data it is (e.g., raw score, standard score, developmental scale score). It includes all variable names and descriptive variable labels. If the variable is categorical, all the possible options should be detailed in the data dictionary (e.g., 1 = *no degree*, 2 = *high school degree*). As previously mentioned, having a systematic way of naming variables is also encouraged (see Reynolds & Schatschneider, 2020, for a discussion of variable naming conventions). If the variable is the transformation of another variable or had been recoded or reverse scored, it should also be documented in the data dictionary. This document can also contain the range of possible scores, the missing values, and/or the amount of complete data. Examples of data dictionaries can be found here: https://www.usgs.gov/products/data-and-tools/data-management/data-dictionaries.

Final Considerations. Practically, these documents can be stored in any digital form (e.g., Microsoft Word). When you are ready to deposit your data, the documentation files can be combined into one, or they can be shared separately for more complex projects. When projects include multiple datasets (e.g., one file for teacher data, one for students), consider providing separate documentation for each dataset.

As a data depositor, it is impossible to predict what a future data user might want to know, so we recommend you provide as much detail as possible about how the data were generated. The time taken to do this will help not only other people use your data, but we promise that you yourself will forget the details of your project over time, and you will thank yourself when you need to reuse your own data.

Deciding Where to Share Data

There are several data repositories that already exist. There are general repositories that contain data from many different disciplines, such as Dataverse (https://dataverse.org/; see Harvard's Dataverse, https://dataverse.harvard.edu/), figshare (www.figshare.com), and Inter-university Consortium for Political and Social Research (www.icpsr.umich.edu). See also re3data, which will let you search for data repositories by keywords and topics (https://www.re3data.org/). Additionally, there are discipline-specific repositories, such as LDbase.org (for educational and developmental science data), and Databrary (www.nyu.databrary.org), a repository for developmental science video data. And finally, some grant funding agencies have their own repositories. For example, the Eunice Kennedy Shriver National Institute of Child Health and Human Development has DASH (https://dash.nichd.nih.gov) and the National Database for Autism Research (https://nda.nih.gov). Each of these repositories, and more, will store your data, and each comes with different features. In general, you should look for a data repository that provides you with a DOI and information about copyright. After that, choosing a repository is a personal preference. This step can be done at any point prior to actual data sharing.

Data Are Ready: Upload

When you are ready to share your data, you have selected your repository; you proceed to the final step, which is uploading your data. Depending on the repository you have picked, the steps will be different. Part of the process will likely involve you selecting the level of access you wish for your data, such as open or restricted access. Once the data are uploaded, the data and the corresponding metadata will be available for others, including yourself, to cite and work from.

Lingering Concerns

As we noted in the introduction, we have some experience navigating the concerns that education scientists have around data sharing. In this section, we summarize some of the most common concerns we have heard, along with our solutions.

"My IRB Will Have a Problem With Me Sharing Data"

Often, investigators are concerned that their IRB will not allow for them to share data already collected without the initial intention to share. This concern is primarily unwarranted, as most IRBs do not have any regulation over data monitoring or data sharing (Bankert & Amdur, 2000). The guidelines for Projection of Human Subjects explicitly states that IRBs should only consider risks directly related to the study (Burnham, 2014), and they are explicitly told to not consider potential (unknown) future risks, a category data sharing can fall into. Additionally, given that most federal agencies require that investigators release their data publicly, IRBs are often receptive to open data practices. Given this, our response to this concern is to recommend talking to your IRB rather than simply assuming that they will say no (see also Meyer, 2018).

Your IRB will likely encourage you to examine your informed consent for guidance about data sharing. If participants consented to their data to remain private and confidential, this does not actually preclude data sharing, because data can be shared after they have been thoroughly de-identified. As mentioned before, if your informed consent language is very restrictive, you should talk with your IRB about a waiver of consent to share de-identified data, likely something they are open to do. In rare cases where discussions with your IRB are not successful, you can still choose to share summary statistics, including metadata, a correlation table, and descriptive statistics. Furthermore, you can report such statistics by key subgroups. New methodologies allow for sample statistics to be of use in modeling (meta-SEM; Becker, 2009), allowing for the data to still be useful for further exploration.

"It Might Be Possible to Identify a Participant"

Many investigators are rightly concerned about privacy issues with sharing data, mostly related to the possibility that it may be possible to identify a given participant. A motivated user could potentially un-anonymize data (e.g., Gymrek et al., 2013). This concern may be magnified because much of educational data are collected on children. However, other than for rare disabilities, the data are likely not overly identifiable. Furthermore, if you use the techniques described in the previous section to anonymize your data, you will be minimizing the potential risk of identifying participants using state-of-the-art de-identifying procedures, while maintaining the known benefits of data sharing practices.

"Preparing and Sharing Data Is a Lot of Work, and Good Help Is Expensive!"

Yes, you are correct. However, funding agencies are aware of this, and therefore, these are allowable costs on your grants. A recent article suggests that 5% to 10% of the budget of a research grant should be dedicated to preparing the collected data to be shared (Mons, 2020). One possible way to structure this is by bringing on data experts, scientists with expertise in data sharing and data management, as collaborators on your grant funded work. Many methodologists or applied statisticians can bring this expertise. Librarians are also trained in data sharing and data management, and most institutions now have scholarly communication, digital scholarship, and data services librarians, who are trained to be part of the research process with the academics at their university. Ideally, such a collaboration would begin before the project starts, with an expert guiding database design and construction with the sharable end product in mind. In this case, the collaboration can be ongoing throughout the life of the project, and so the study documentation can be built slowly over the course of the project as decisions are made. Preparing data for sharing can be done after the study is completed, but it requires more dedicated time to remember and reconstruct details of data collection processes and decisions. Either way, the 5% to 10% budget benchmark is a great way to keep data reuse in mind as you plan a budget for a new project. While it may be slightly more expensive for you right now, it will be less expensive to the world in the long term.

"What If I Share My Data and Someone Scoops Me?"

This concern less colloquially can be articulated as a worry that someone will publish your key findings or specific aims from your project before you are able to do so. We have a few responses to this. First, the timeline of data sharing activities is not mandated. As we previously noted, it is not typically expected that data will be shared immediately after collection; datasets are more commonly shared only after the primary aims have been published. If you have other work you want to do with your data, many data repositories have an embargo feature, which allows you to put your data online but have it masked from public view for a given period of time (i.e., 1–4 years). However, there is little evidence of this worry manifesting in the academic space.

"No One Will Collect Data If All of the Data Were Open and Free"

This is a slippery slope argument that on examination falls apart. First, research is often dependent on particular measures, and sometimes, those are newly developed measures. Novel measures or measurement strategies will always need to be collected. Second, much of education

science works in the area of intervention science. People want to test the efficacy or effectiveness of their interventions on a particular sample. Even if data were widely available and freely open, researchers would still be developing new technologies and new areas in which to implement them. We will always be in the business of figuring out if a particular intervention works on a particular skill in a particular sample of participants, and therefore, we will almost certainly always be involved in collecting the data that suit those purposes.

"Data Sharing Is Fine for Others, but My Situation Is Different and I Don't Need to Share My Data"

This is a classic case of exceptionalism thinking, and may be a common response, but in reality, there are very few cases where sharing data would not be a beneficial or worthwhile activity. With primary data, it might seem like you or your research team are the only ones who could benefit or use the data you have collected, but it is not always obvious how the data you collect may be useful to another scientist. We assure you that your data or your situation is almost certainly not different; every project, dataset, and situation is unique in its own way.

"I am Not Able to Share My Data"

In some cases, an educational researcher has a legitimate reason for not being able to openly share their data. For example, some investigators might receive progress monitoring data from their partner school district, which they are expressly forbidden from sharing. IES lists the example of proprietary data as a different data source that might not be shareable (https://ies.ed.gov/funding/datasharing_implementation.asp). In these cases, we suggest that researchers consider if they can share some of the data, despite other parts of the data not being shareable. With the data they cannot openly share, researchers can consider determining if strict data use agreements might allow some sharing to occur with specific researchers, which opens a path for others to use the data, albeit with more effort. Researchers can also consider if they are able to share their data in a repository that allows access restriction (e.g., protected access, see list here: https://osf.io/tvyxz/wiki/8.%20Approved%20Protected%20Access%20Repositories/). Such restricted access still allows metadata to be available to others, and there is a clear way for data access to be requested. Finally, you can still participate in data sharing by publishing the processed data (the means, standard deviations, and correlation matrices used in your study), or the code you used to create or recode the key variables or select the cases. Any of these methods are preferable to "data available upon request" or simply not available at all.

"These Data Exist Because of My Hard Work, Why Should Someone Else Benefit?"

This particular concern is personal and is related to simply not wanting to give data to others. Likely this is a product of their cultural academic upbringing. Because it is relatively ingrained in scholars' identities, it is a particularly difficult concern to overcome, and it is likely the most pervasive in the field (see also Cook et al., 2018, for similar calls in special education). If this message resonates with you, or with a colleague whom you are working with, we have the following suggestions. First, we hope that the advantages that we have documented in the earlier sections of this article will help guide this discussion. In addition, remember that data sharing is not all or none. Consider sharing the data you have already published on. Or consider preparing your data for sharing and posting it to a repository, but allowing reuse through application only or after a certain amount of time (these are options available with some repositories, such as LDbase.org).

Conclusion

In this article, we describe the importance of data sharing and the steps to take to prepare your data for sharing. There are many good reasons to share your data, for yourself, and for the educational sciences community and stakeholders. Properly sharing your data is not a small task, but it is not impossible and, if planned, can be budgeted and the work spread across the project timeline. Through federal mandates, followed by community buy-in, data sharing is becoming more and more accepted and expected. We anticipate that all educational sciences investigators should prepare for data sharing to be part of their research process, and we urge our colleagues to remember that there are many ways to share your data and that sharing some data is better than sharing no data.

Acknowledgments

This work is supported by Eunice Kennedy Shriver National Institute of Child Health and Human Development Grants HD052120 and HD095193. Views expressed herein are those of the authors and have been neither reviewed nor approved by the granting agencies.

ORCID iD

Jessica A. R. Logan https://orcid.org/0000-0003-3113-4346

Notes

1. We differentiate community-led data sharing from the large relatively openly available datasets that were purposely collected to be shared, such as the National Longitudinal Surveys or the Early Childhood Longitudinal Survey. These purposefully collected datasets are incredible resources for the educational sciences

community and thus are great examples of the potential of data reuse, but they are rare and do not represent the situation most educational researchers find themselves in when considering sharing their own data.

2. This can be done through a "save as" feature, or the export feature of most statistical programs with an interactive graphical user interface (pull-down menus). Note that in some programs, you will have to select a type of encoding, and we suggest the Unicode option.

3. An example of informed consent text that clearly describes your intention to openly share your data, in the experience of some of the authors:

> Original records and identifiable data will be heard or viewed only for research purposes by the investigator and his or her associates. Data with all identifiers removed may be used for future projects that focus on any topic and may be unrelated to this study. This new data may be made available to the general public via the internet and an open database. This information will not have your name or other personally identifiable information included (i.e., it will be de-identified).

See Shero and Hart (2020a) for more suggestions.

4. We strongly urge researchers to consider sharing those old datasets you have sitting around that you are done with; they might be exactly what someone else needs for their research questions.

References

Albro, E. (2020, January). *IES annual principal investigators meeting*. https://ies.ed.gov/pimeeting/

Bankert, E., & Amdur, R. (2000). The IRB is not a data and safety monitoring board. *IRB: Ethics & Human Research*, *22*(6), 9–11. https://doi.org/10.2307/3563586

Becker, B. J. (2009). Model-based meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis* (pp. 377–395). Russell Sage Foundation.

Burchinal, M., & Neebe, E. (2006). Best practices in quantitative methods for developmentalists: I. Data management: Recommended practices. *Monographs of the Society for Research in Child Development*, *71*(3), 9–23. https://doi.org/10.1111/j.1540-5834.2006.00354.x

Burnham, B. (2014). Open data and IRBs. *Open Science Collaboration*. http://osc.centerforopenscience.org/author/bryan-burnham.html

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PLoS One*, *15*(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

Cook, B. G., Lloyd, J. W., Mellor, D., Nosek, B. A., & Therrien, W. J. (2018). Promoting open science to increase the trustworthiness of evidence in special education. *Exceptional Children*, *85*(1), 104–118. https://doi.org/10.1177/0014402918793138

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

Drachen, T. M., Ellegaard, O., Larsen, A. V., & Dorch, S. B. F. (2016). Sharing data increases citations. *LIBER Quarterly*, *26*(2), 67–82. https://doi.org/10.18352/lq.10149

Edwards, A., & Schatschneider, C. (2020). *De-identification guide* (Version 1). figshare. https://doi.org/10.6084/m9.figshare.13228664.v1

Ember, C., Hanisch, R., Alter, G., Berman, H., Hedstrom, M., & Vardigan, M. (2013). *Sustaining domain repositories for digital data* [White paper]. https://deepblue.lib.umich.edu/handle/2027.42/136145

Federal Committee on Statistical Methodology. (2005). *Statistical policy: Report on statistical disclosure limitation methodology* (Working Paper 22; 2nd version). https://www.hhs.gov/sites/default/files/spwp22.pdf

Freese, J. (2007). Replication standards for quantitative social science: Why not sociology? *Sociological Methods & Research*, *36*(2), 153–172. https://doi.org/10.1177/0049124107306659

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, *339*(6117), 321–324. https://doi.org/10.1126/science.1229566

Khan, K. S., Purtell, K. M., Logan, J., Ansari, A., & Justice, L. M. (2017). Association between television viewing and parent-child reading in the early home environment. *Journal of Developmental & Behavioral Pediatrics*, *38*(7), 521–527. https://doi.org/10.1097/DBP.0000000000000465

Langille, M. G., Ravel, J., & Fricke, W. F. (2018). "Available upon request": Not good enough for microbiome data! *Microbiome*, *6*, Article 8. https://doi.org/10.1186/s40168-017-0394-z

Logan, J. (2019). *Data management and data management plans* [PowerPoint slides]. figshare. https://doi.org/10.6084/m9.figshare.7890827.v1

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Makel, M. C., Hodges, J., Cook, B., & Plucker, J. (2019, October 31). *Both questionable and open research practices are prevalent in education research*. https://doi.org/10.35542/osf.io/f7srb

Mannheimer, S., Pienta, A., Kirilova, D., Elman, C., & Wutich, A. (2019). Qualitative data sharing: Data repositories and academic libraries as key partners in addressing challenges. *American Behavioral Scientist*, *63*(5), 643–664. https://doi.org/10.1177/0002764218784991

McBee, M. T., Makel, M. C., Peters, S. J., & Matthews, M. S. (2018). A call for open science in giftedness research. *Gifted Child Quarterly, 62*(4), 374–388.

Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in Methods and Practices in Psychological Science*, *1*(1), 131–144. https://doi.org/10.1177/2515245917747656

Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature*, *578*(7796), 491–491. https://doi.org/10.1038/d41586-020-00505-7

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, Article 21. https://doi.org/10.1038/s41562-016-0021

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS One*, *2*(3), e308. https://doi.org/10.1371/journal.pone.0000308

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, Article e175. https://doi.org/10.7717/peerj.175

Reynolds, T., & Schatschneider, C. (2020). *The basics of data management*. figshare. https://doi.org/10.6084/m9.figshare.13215350.v1

Shero, J., & Hart, S. A. (2020a). *Informed consent template* (Version 1). figshare. https://doi.org/10.6084/m9.figshare.13218773.v1

Shero, J., & Hart, S. A. (2020b): *IRB protocol template* (Version 1). figshare. https://doi.org/10.6084/m9.figshare.13218797.v1

Shero, J., & Hart, S. A. (2020c). *Working with your IRB: Obtaining consent for open data sharing through consent forms and data use agreements* (Version 1). figshare. https://doi.org/10.6084/m9.figshare.13215305.v1

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., . . . Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Tsai, A. C., Kohrt, B. A., Matthews, L. T., Betancourt, T. S., Lee, J. K., Papachristos, A. V., Weiser, S. D., & Dworkin, S. L. (2016). Promises and pitfalls of data sharing in qualitative research. *Social Science & Medicine*, *169*(November), 191–198. https://doi.org/10.1016/j.socscimed.2016.08.004

U.S. Department of Health & Human Services. (2012). *Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule*. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

U.S. Department of Health & Human Services. (1974). *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J.-S., Renaut, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, *24*(1), 94–97. https://doi.org/10.1016/j.cub.2013.11.014

Vision, T. J. (2010). Open data and the social contract of scientific publishing. *BioScience*, *60*(5), 330–331. https://doi.org/10.1525/bio.2010.60.5.2

The White House. (2013, May 9). *Executive order: Making open and machine readable the new default for government information*. https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728. https://doi.org/10.1037/0003-066X.61.7.726

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Bouwman, J. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Authors

JESSICA A. R. LOGAN is an assistant professor of educational studies at The Ohio State University. She works in the field of quantitative methodology and statistics, identifying new statistical models or research designs and adapting them for questions about how children grow and change, as well as supporting researchers to collect, manage, and share high-quality data.

SARA A. HART is an associate professor of psychology at Florida State University and the Florida Center for Reading Research. Her substantive research relates to understanding how and why people differ in their cognitive development, particularly focused on reading and math development, and she is also interested in supporting rigorous and reproducible educational and developmental science.

CHRISTOPHER SCHATSCHNEIDER is a professor of psychology at Florida State University and the Florida Center for Reading Research. His interests are in the identification of skills that are related to the acquisition of reading ability and the use of these skills to identify children who are at risk for early reading problems, as well as in research design, measurement, and statistical methodology.