Institute of Education Sciences

Toolkit

Sharing Study Data: A Guide for Education Researchers

NCEE 2022-004
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation and Regional Assistance

The Institute of Education Sciences (IES) is the independent, non-partisan statistics, research, and evaluation arm of the U.S. Department of Education. The IES mission is to provide scientific evidence on which to ground education practice and policy and to share this information in formats that are useful and accessible to educators, parents, policymakers, researchers, and the public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other IES product or report, we would like to hear from you. Please direct your comments to ncee.feedback@ed.gov.

This report was prepared for the Institute of Education Sciences (IES) under Contract 91990020F0052 by Mathematica. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

March 2022

IES Institute of Education Sciences

Sharing Study Data: A Guide for Education Researchers

Ruth Curran Neild
Danielle Robinson
Jacqueline Agufa
Mathematica

NCEE 2022-004
U.S. DEPARTMENT OF EDUCATION

Contents

Sharing Study Data: A Guide for Education Researchers

Open science envisions that researchers will make their study data available to other investigators to facilitate research transparency and accelerate the development of knowledge. This guide describes key issues that education researchers should consider when deciding which study data to share, how to organize the data, what documentation to include, and where to share their final dataset. The guide also provides strategies for addressing related challenges and includes links to other resources, a checklist aligned to each section, and appendices that contain templates and samples.

Sharing research findings is at the heart of the scientific enterprise. Education researchers routinely share their results in published studies for others to examine, debate, build on, and seek to replicate. Over the past decade, the education research community has broadened access to these findings by making more research freely available through open-access journals, research collections like ERIC, or university websites. This open science approach, also known as open access, envisions that researchers will also make their study data available to other investigators to facilitate research transparency and accelerate the development of knowledge.

Data sharing is not new to the education research field. For example, the Inter-university Consortium for Political and Social Research (ICPSR) was organized in 1962 and still serves as a key repository for data from completed studies in education and other social science fields. What *is* new is a 2013 requirement for open access to digital data from federally-funded scientific studies (see Box 1 for more information about this requirement and for links to agency policies). This requirement, along with growing interest in open science practices, means that many more education researchers are looking for practical, achievable strategies for making their data available to others.

Why this guide

Making data available for broader use involves making decisions and using procedures that may be unfamiliar to many education researchers. Creating a dataset that another investigator can use independently, without support from the original study team, requires special attention to data organization and documentation. Sharing data from education studies can also present complex challenges related to participant confidentiality and data ownership. Researchers are responsible for taking reasonable steps to reduce the risk of disclosing the identity of the individuals or entities in their studies who were promised confidentiality. Further, some data may be owned by an education agency that has, in effect, loaned the data for that particular study. To share the data for use beyond the study, the researcher must ensure that they have permission from the agency.

Box 1: Federal policies on open access to research and data

In 2013, the U.S. federal government announced a commitment to ensuring public access to scientific publications and digital data funded in whole or in part with federal funds (Holdren, 2013). Each agency with relevant scientific activities was tasked with developing its own policy. Federally-funded researchers should become acquainted with the specific policies of their funding agency. For education researchers, the policies of greatest interest are likely to be those of the following agencies:

- The U.S. Department of Education, Institute of Education Sciences
- The National Institutes of Health
- The National Science Foundation

This guide describes key issues that education researchers should consider when deciding which study data to share, how to organize the data, what documentation to include, and where to share their final dataset. The guide also provides strategies for addressing those challenges and includes links to other resources, a checklist aligned to each section, and appendices that contain templates and samples. The content is most relevant to investigators sharing quantitative data rather than text, audio, or video files, although the guide provides links to resources on sharing these kinds of data. Sharing qualitative data presents special challenges for maintaining confidentiality, and standards and best practices are emergent.

This guide is intended as a starting place–an opportunity to become acquainted with the basics of data sharing and prepared to engage additional expertise as needed. The guide provides a way of thinking through typical challenges associated with sharing data and a set of tools researchers can use to develop and implement their plans.

Who should use this guide

The guide is designed for education researchers who are planning to share their study data or are considering whether to do so. It describes steps researchers can take at the beginning, middle, and end of their studies. It also includes suggestions for researchers who are nearing the end of their projects without having decided on an approach to data sharing. Researchers who are preparing applications for funding for new studies can also consult this guide for examples of sharing study data under different circumstances and to understand the resources that might be required to prepare a complete dataset.

Box 2: IES resources on public access to data

IES Policy Regarding Public Access to Data Resulting from IES Funded Grants

Implementation Guide for Public Access to Research Data

Frequently Asked Questions About Providing Public Access to Data

A key audience for the guide is investigators with research grants from the U.S. Department of Education's Institute of Education Sciences (IES) that require sharing data from completed studies. The guide complements, but does not replace, IES policy statements and guidance about public access to research data (see Box 2 for links to IES data sharing requirements and guidance). The guide's recommendations for disclosure risk management, data file organization, and data documentation are consistent with IES's approach to sharing restricted-use data from national education evaluations that it conducts.

How this guide is organized

The guide is organized around three key issues that researchers should consider as they plan and execute their approach to sharing data. We characterize these issues as the "three Ds" of data sharing: disclosure risk management, data documentation and organization, and depositing the dataset. Each issue is the focus of a separate section in this guide, as follows:

- **Section I: Disclosure risk management.** This section addresses appropriate authorizations to share all or part of the study data and describes strategies to reduce the risk of disclosing the identities of individuals and other entities, such as schools or districts, who were promised or can reasonably expect confidentiality.
- **Section II: Documentation and organization.** This section illustrates how data files, documentation, and code can be organized to allow another investigator to understand the dataset and conduct analysis using the data files.

- **Section III: Depositing the dataset.** This section presents considerations about where to house the dataset, given its content, the requirements of any outside data providers, and the importance of informing other researchers that the dataset is available.

Researchers who are starting a new study may need to develop a Data Management Plan (DMP) and share it with their funder to document their anticipated approach to data sharing. Each of the three sections includes information to support developing a DMP (Box 3). Regardless of whether a DMP is required by a funder, the elements of the DMP are important considerations for researchers who are in the planning phase of their study.

Box 3: Using this guide to develop a Data Management Plan for a new study

A Data Management Plan (DMP), prepared at the start of a study, describes how the researcher plans to share their study data. Some funders, including IES, require grant-funded researchers to provide a DMP upon award. Various tools provide guidance on constructing a DMP. Together, the three sections of this guide address key information that a DMP should cover, including:

- Types of data to be shared
- Data sharing agreements
- Procedures for managing disclosure risk
- Where the data will be shared
- Documentation to be provided

The ordering of the three sections–with disclosure risk addressed first–reflects a logical flow of contingencies from determining latitude to share data to choosing how and where to share the data. However, for a study just getting underway, developing good procedures for data documentation and organization (Section II) should be one of the first tasks the study team undertakes. Further, the issues discussed in the guide are interconnected, and decisions related to one issue may affect options for the others. Table 1 summarizes key steps related to sharing data across a study lifecycle and points to the section of this guide where each step is discussed.

Table 1: Steps to prepare data for sharing occur across the study lifecycle

Planning the study	Beginning the study	During the study	Concluding the study
Discuss data sharing plans/requirements with potential partner schools/agencies (Section I) In project budgets, include costs for data management across the lifetime of the study and preparation of a dataset for sharing (Section II)	Develop consent forms and memoranda of understanding that allow for data sharing (Section I) Develop procedures and practices for organizing study data and managing code, potentially engaging database expertise to do so (Section II)	Maintain ongoing documentation about project measures, data collections, and decisions (Section II) Begin to explore data repositories for where study data will be shared (Section III)	Examine data for direct and indirect identifiers (Section I) Explore mitigation strategies and assess their feasibility and usefulness for data sharing (Section I) To create the final dataset, include a Description of Data Files and relevant code, drawing on records kept and code developed by the team throughout the study (Section II) Finalize choice of data repository (Section III)

Keeping the big picture in mind

- Creating datasets that are useful and accessible to other investigators and properly manage disclosure risk requires attention to many details. However, even when working through the details, it can be helpful to keep the purpose and benefits of data sharing in mind. The following perspectives inform each section of the guide: **The goal of data sharing is to produce something of value for science and, ultimately, for the improvement of education.** In the spirit of open science, researchers should take care to provide well-organized, well-documented data files that other investigators can use with a modest investment of effort. Including a range of variables beyond those used to produce the study's main findings can add value by enabling another investigator to examine mediators, moderators, and other outcomes.
- **Focusing on sharing a well-organized and well-documented dataset can improve the organization and efficiency of the original study team.** Carefully recording study decisions and data collection during the course of the study and developing routines for organizing code and data files can improve the study team's own work by reducing confusion, duplication, and inaccuracies.
- **Researchers should commit to sharing some data or code to facilitate additional analysis.** Even if there are limited options for sharing all or part of the original study data, researchers should consider what else they can share, such as code or aggregated data.
- **There is no single approach to data sharing.** Each study has a unique set of opportunities and constraints. Learning how other researchers have approached data sharing can provide ideas that might work for a given study. However, it might not be possible to use exactly the same approach that another researcher has used.
- **Tradeoffs may be necessary.** Researchers will need to find a defensible balance of disclosure risk, completeness of the data, and broad access to the dataset. For example, a complete dataset could enable the reproducing of study results but involve more restrictions on who can access the data. Alternatively, a partial dataset might allow more access to the data but only be useful for limited new analyses.

Section I: Disclosure Risk Management

Disclosure risk management is a consideration for education researchers even when they do not plan to share data outside their research teams. For example, researchers commonly pledge to maintain the confidentiality of their study participants' data and follow data management processes to reduce the risk of disclosing personally identifiable information (PII) (see Box 4 for definitions of PII, confidentiality, and disclosure risk management). If the study uses student data obtained from a state or local education agency, the researcher would have signed an agreement committing to use the data in accordance with federal requirements specified in the Family Educational Rights and Privacy Act (FERPA) and any applicable state and local regulations.

Managing disclosure risk in shared datasets builds on the practices that education researchers are expected to use in their own teams to mitigate disclosure risk. Although there may be unusual cases where sharing study data is inadvisable because the risk and potential harm of disclosure is so great, there are many different tools that can strengthen protections against unauthorized disclosure and make data sharing possible.

Box 4: Key definitions

- **Personally identifiable information** is information that can be used to identify an individual or trace their identity when used alone or in combination with other information (Johnson 2007).
- **Confidentiality** means that the participant's PII is not disclosed to unauthorized persons (NCES 2011).
- **Disclosure risk management** is the practice of assessing potential threats to participant confidentiality and taking appropriate steps to reduce or eliminate these risks.

All data sharing involves disclosure risk. Even the simplest dataset—for example, one with just a few variables and no direct or indirect identifiers—has the potential to be combined with other information, now or in the future, that could identify study participants. The risk of disclosure is never zero. However, the risk can be reduced when researchers take reasonable steps to mask participants' identities and, if necessary, restrict access to the data. Researchers planning a data-sharing strategy should focus on taking reasonable steps to eliminate obvious sources of disclosure risk.

This section addresses researchers' authorization to share data and offers brief descriptions of specific, achievable strategies for creating datasets that can be shared with other investigators while maintaining participant confidentiality. Links to additional resources provide more in-depth guidance on how to implement specific disclosure risk mitigation strategies.

Ensuring authorization to share data

Disclosure risk management begins with an assessment of whether the researcher has obtained appropriate authorizations from data owners and study participants that allow the

data to be shared beyond the study team. Researchers should work closely with their Institutional Review Board to determine their latitude to share data from specific studies.

Researchers' latitude depends, in part, on who owns the study data and the owner's willingness to allow the data to be shared beyond the original study team. Researchers whose studies incorporate non-public data from local or state education agencies or postsecondary institutions must be certain that they have proper authorization from these data owners before sharing any data. They must also understand clearly the format in which the data owner will allow data to be shared (for example, as individual-level data or only in the aggregate). In addition, they should be sure that there is no language in participant consent forms that explicitly prohibits data sharing.

Researchers with studies already underway should read any data agreements carefully, discuss options transparently with the agencies or institutions providing data for the study, and consult with their Institutional Review Board. Researchers planning new studies or just beginning their projects should be straightforward with partner agencies or institutions about their plans to share data and funders' data sharing expectations.

New data agreements, as well as participant consent forms, should explain clearly and simply how data will be used, including the possibility of data being shared. ICPSR offers a resource on writing consent forms and interpreting consent forms that were developed without data sharing in mind (see Box 5).

Box 5: Resources on participant consent and data sharing

The Inter-university Consortium on Political and Social Research (ICPSR) provides specific suggestions for language to include in consent forms that allows for data sharing. At this link, ICPSR also addresses concerns that researchers may have about sharing data when their consent forms were not specifically developed with data sharing in mind.

LDBase, an education data repository, also offers a resource with suggestions for consent form language.

Assessing disclosure risk in data files

Assessing disclosure risk requires looking carefully at the information in the data files and understanding the risks that may be unique to those files. Researchers planning to share both raw and analytic files will need to conduct a disclosure risk analysis for both types of files (see Box 6 for more information on file types). The disclosure risk analysis is a systematic assessment of whether variables in the data file contain PII that could individually or in combination identify a study participant or pose a special concern because of their sensitive nature. Mapping these variables using the steps described next will help clarify which approaches to mitigating disclosure risk might be needed.

Step 1: List variables in the data files that directly identify an individual

Direct identifiers contain information unique to the individual that can be used to directly link them to their data. Direct identifiers will almost always need to be deleted from the shared data file. In education research, examples of these variables include student name, names of parents or other family members, address of the student or their family, Social Security number, or an individual identification number used by the school or district. If a district or school was promised confidentiality, variables directly identifying these entities should be flagged as well. If the original research team is following good disclosure risk management practices, most or all of this information about individuals, schools, or districts would have been removed when creating its own analytic files.

Step 2: List potential indirect identifiers

Indirect identifiers are variables that could be used in combination with each other or with other extant data sources to identify an individual. Indirect identifiers generally are demographic or other characteristics that could reasonably be observed by someone in the school or community or knowable from public data sources. Examples of these variables could include birth date, place of birth, gender, race or ethnic background, special education classification, or geography. Keep in mind that indirect identifiers can be variables at the class, grade, school, or district levels, not just at the individual level. In addition, in education contexts, information potentially known to school administrators, teachers, or parents and friends of students should be considered potential indirect identifiers.

Box 6: Analytic and raw (source) data files

At a minimum, researchers should plan to share their analytic data files, but raw, or source, data files can also be useful to other investigators. IES encourages its grantees covered by the IES Policy Regarding Public Access to Data Resulting from IES Funded Grants to provide the source data for any constructed variables.

- **Analytic data files** are created by the researcher for the analysis reported in the study manuscript. They include data for individuals included in the analysis and the constructed variables.
- **Raw, or source, data files** contain the original data prior to preparation for analysis. For example, a raw file might include the variables that were combined to create a constructed variable, indicate missing data, and include the full sample before inclusion rules were applied. Basic cleaning may have been carried out on these data.

Raw data files provide for maximum transparency, consistent with the aims of public access, because they provide the best visibility into the analytic choices made by the original research team. Some data owners, such as education agencies, may be less willing to share raw data when the data are more granular than the analytic data. We recommend providing whatever raw data can be shared, even if just a subset of the raw data that was used for the study.

Even if a school or district is not named, data taken from public sources, such as school enrollment or the percentage of students from low-income households, could be used in combination to determine its identity. The more granular the data—for example, the more detailed a student's disability classification or the more precise the number of students enrolled in each grade—the easier it will be to use these variables alone or in combination to identify student participants.

As part of this exercise, consider whether any published manuscripts or presentations from the study, or even media attention about the study, include information that could directly or indirectly identify the schools or districts where the study was conducted. Even if the schools or districts have not sought confidentiality, the fact that this information has been disclosed requires extra attention to the risk of disclosing individual identities.

Step 3: Examine cell sizes of indirect identifiers (individually and in combination), and flag small cell sizes for disclosure risk mitigation

Individually or in combination, some variables may be able to indirectly identify individuals in the data file when there are few people with specific groupings of those characteristics. For example, a user might be able to identify both a school and a specific teacher in that school by combining four variables from a study data file (school enrollment, subject taught, gender, and race/ethnicity) with public school-level data sources. Adding another variable, such as a specific learning disability, might enable identification of a specific student.

Research teams can use simple frequencies and crosstabulations to assess the potential of indirect identifiers, individually or in combination, to disclose participant identity. We recommend adhering to the National Center for Education Statistics (2012) statistical standard 4-2-10, which specifies that cells produced by combining indirect identifiers contain at least three cases. Variables with three or fewer cases can be modified to be less granular or, in some cases, might need to be deleted from the dataset.

Step 4: Identify variables that contain information of a sensitive nature

Some types of information could cause special harm or embarrassment to individuals if disclosed. In education research, examples of such information might include specific special education classifications or disciplinary histories. These data do not necessarily need to be removed from the dataset, but they may warrant stronger risk mitigation strategies.

Mitigating disclosure risk in data files

Having examined the data files for direct identifiers, indirect identifiers, and variables with sensitive information, the next step is to assess options for mitigating the risks. For direct identifiers, the solution is clear: delete them from the data files or anonymize them with pseudonyms, such as identification numbers that consist of random character strings. To mitigate risks of indirect identifiers and sensitive variables in the data files, two types of solutions can be used separately or in combination: (1) technical solutions, which make

changes to the data file to obscure individual identities (also known as statistical disclosure control) and (2) procedural solutions, which limit who can access the data and the environment in which they can access it.

We recommend first exploring the feasibility of technical solutions, followed by procedural solutions if needed. A key part of this exploration is to assess how much the dataset would be degraded—that is, become notably less useful for reproducing or extending study findings—by technical solutions. If the value of the dataset for transparency and knowledge development would be undermined in important ways, providing restricted access to a more complete dataset is likely the preferred way to meet the aims of open science. This decision may require a judgment call by the researcher about the utility of the resulting data.

Step 1: Assess feasibility of technical solutions to mitigate disclosure risk, and examine their impact on the usefulness of the dataset

Technical solutions for mitigating disclosure risk range from basic changes to variables and file structure to advanced approaches that require special expertise or tools to execute. Basic changes include deleting variables, recoding variables to make them less granular, or creating multiple files that cannot be linked to each other and thus prevent disclosure through combinations of indirect identifiers. Advanced approaches include data swapping, micro-aggregation, and data perturbing. Each of these approaches is discussed next, and an annotated bibliography at the end of the section provides links to additional resources.

In selecting an approach, researchers should assess how modifying the data would affect the ability of another investigator to reproduce the analysis as reported in the study or to use the data for secondary analysis. For example, if the study uses an outcome variable that is continuous, consider whether recoding the variable to make it less granular will render it impossible for another investigator to reproduce or closely approximate the study's findings.

Researchers should consider the approach that their resources—expertise, time, and project budget—will allow. As they plan budgets for new studies, researchers might want to consider including resources for consultation with disclosure risk experts or consideration of published work by experts to prepare files for data sharing. Several data repositories developed or commonly used by education researchers provide data consultation and curation services (see Section III).

Basic changes to variables and file structure

Basic changes to variables and file structure might need to be implemented and assessed iteratively to understand their value for mitigating disclosure risk without substantially degrading the usefulness of the dataset. Types of basic changes to the data include:

- **Assigning new, study-specific, unique identifiers, ensuring random order.** To ensure that the identifiers for individuals or clusters such as schools or districts do not

reflect alphabetization or any other ordering that could be used to infer identities, randomly order observations in each level of the dataset before assigning identifiers.

- **Coarsening variables that could indirectly identify participants.** Coarsening reduces the amount of detail a variable provides by combining values into broader intervals, which can help mitigate disclosure risk associated with small cell sizes. The data provided remain accurate, although less precise. Techniques include:
    - *Recoding quantitative data into categories.* For example, rather than providing a specific age, a dataset could provide age ranges. For school- or district-level variables obtained from publicly available sources, such as enrollment or the percentage of students receiving free or reduced-price lunch, recoding into categories can help obscure the location where the study was conducted.
    - *Top and bottom coding*. Values at the top and bottom of a distribution can be combined into a single category. For example, income could be capped with an average of the values over a given amount.
    - *Recoding categorical data into broader intervals.* For example, smaller geographic units could be combined into a larger unit.
- **Recoding open-ended questions into categorical variables**. Code responses to open-ended questions into broad categories, or delete them from the data file if recoding is not possible or not meaningful.
- **Creating a set of unlinked files.** If the preceding basic strategies do not sufficiently reduce the risk of disclosure while also preserving information needed for analysis, another option is to create multiple data files, each with a subset of the variables that allow a specific analysis from the study to be reproduced. The files must not include variables that would allow the data files to be combined. A disadvantage of using unlinked files is that while they can support reproducing the findings from a study, they might hamper opportunities to conduct new analyses of connections among variables that appear only in one data file or another.

Advanced technical solutions for mitigating disclosure risk

Advanced techniques for mitigating disclosure risk have been used to develop public use data sets, such as those available from the U.S. Census or the National Center for Education Statistics at IES. These techniques slightly distort the data to make it more difficult to identify an individual or entity with certainty. Advanced techniques include:

- **Data swapping**. This strategy targets specific variables in a dataset, exchanging actual values between sets of individuals who are the same or similar on other key variables. For example, for two individuals from different cities who have the same age, gender, and race/ethnicity, uncertainty about their identities could be introduced by assigning each one the other's geography. A sample of cases are swapped in this way, with the percentage of cases swapped held confidentially.

- **Micro-aggregation**. This strategy replaces individual values with average values computed on small aggregates. For example, sets of three adjacent values of a variable requiring special protection could be recoded to their average value. This approach can be used for variables where averaging is meaningful (for example, school enrollment).
- **Perturbing.** This process adds random statistical "noise" to the data so that individual values are changed while the mean of the variable is preserved. The more noise is added, the more the disclosure risk is reduced.

Advanced solutions can be used along with basic changes to the data: for example, variables can first be recoded to coarsen them, and then some values on key variables can be swapped to provide additional protection.

Because these advanced techniques distort the data and add variance to the estimates, it is important to assess their effects on the data. For example, while swapping preserves the univariate distributions for the variables of interest, it can affect correlations between variables. With micro-aggregation, it is important to assess how much the modified data differ from the original data in terms of the distributions of variables and their correlations. Likewise, for perturbing, introducing more noise can make it more difficult to reproduce the findings of the original study.

Implementing these advanced methods is a complex undertaking. For this reason, if these methods are potentially of interest, we recommend consulting with experts, such as those affiliated with data repositories. An alternative is to use statistical code written by experts and designed to implement and assess the impact of these methods on disclosure risk mitigation and data utility (such as an R package developed by Templ, Kowarik, and Meindl (2015), described in the annotated bibliography).

Assessing utility and risk in the altered data

After implementing basic changes or advanced methods to blur the identities of study participants, the next step is to examine the utility of the resulting data. Fundamentally, the question is how different the altered data are from the original data. To assess the differences, researchers should explore the univariate distributions and crosstabulations for pairs of variables. Ideally, the statistics for important variables in the dataset—means, variance, covariance, and correlation—will not differ between the altered data and the original dataset (Benschop & Welch, n.d.). An additional step is to compare the estimates from the altered data against the original study estimates by re-running the models with altered data; if the main findings from the original study cannot be closely reproduced, options include taking another approach to altering the data (for example, adding less "noise") or restricting access to the data (see Step 2). Likewise, researchers should examine whether the altered data mitigates disclosure risk to the extent expected (for example, cell sizes with no fewer than three individuals, per NCES standards described previously).

Implementing changes to the data and assessing their impact on utility and disclosure risk might be an iterative process until a good solution is found—or until it becomes clear that the best option for sharing data involves a procedural solution, described next.

Step 2: If needed, mitigate disclosure risk by restricting access to the data

If making changes to the data as described in Step 1 would substantially degrade the data's utility for research, a procedural solution—restricting access to the data—might be the right approach. Restricted-use datasets place limits on who can access the data and, in some cases, the method by which they can access it. These datasets can be shared through data repositories or archives, through the education agency or institution that provided the data, or directly by the researcher who is sharing the data. Not all data repositories accept restricted-use data (see Section III on data repositories).

Data repositories take different approaches to vetting researchers who apply to use the data: some assess applicants against a set of criteria, while others require the original researcher to grant permission for its use. Some repositories allow researchers to select which of these options they prefer. Prospective data users might be required to present appropriate academic credentials. Investigators using these datasets pledge to maintain study participant confidentiality and to take steps to prevent unauthorized access to the data, and there can be penalties for violating these agreements.

Access to restricted-use data files can be provided with an electronic copy of the data, through a Virtual Data Enclave that allows researchers to conduct analyses of the data without downloading a dataset, or, for very sensitive data, conducting analysis at a specified, on-site location.

To maximize the free flow of information and potential advancement of knowledge, researchers should consider whether it is feasible to share at least some of their data in a public-use format, even if the full dataset might need to be restricted to maintain participant confidentiality. Releasing multiple datasets from a study requires special care so that no inadvertent inconsistencies between the data files are introduced or disclosure risks created.

Additional resources about disclosure risk management

Background on methods for assessing and mitigating disclosure risk

Federal Committee on Statistical Methodology. (2005). Report on statistical disclosure limitation methodology. Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical and Science Policy. Retrieved January 16, 2022 from https://www.hhs.gov/sites/default/files/spwp22.pdf.

This is an exhaustive report prepared by representatives of federal statistical agencies, with additional expert consultation. The report covers a wide range of confidentiality issues. Chapter 5, "Methods for Public-Use Microdata," is likely to be of most interest to individual

researchers. The chapter describes both basic and advanced methods for mitigating disclosure risk and includes links to additional technical resources on these methods.

Garfinkel, Simson L. (2015). De-Identification of personal information. U.S. Department of Commerce, National Institute of Standards and Technology. Retrieved January 16, 2022 from https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf.

This publication, written for a broad audience of policymakers and researchers across a range of disciplines, provides an overview of concepts, terminology, and strategies related to assessing and mitigating disclosure risk in shared data sets. It does not provide detailed how-to information or delve into technical details of mitigation strategies.

How-to guides for researchers

Benschop, T. & Welch, M. (n.d.). Statistical disclosure control: A practice guide. Retrieved January 16, 2022 from https://readthedocs.org/projects/sdcpractice/downloads/pdf/latest/.

Although the primary purpose of this guide is to explain how to use features in the sdcMicro software, described later in this reference list, the guide includes detailed descriptions of disclosure risk assessment and mitigation approaches. The guide also shows how to implement these solutions in the sdcMicro software.

Inter-university Consortium for Political and Social Research. (n.d.) Guide to social science data preparation and archiving: Best practice throughout the data life cycle (6th edition). Retrieved January 16, 2022 from https://www.icpsr.umich.edu/files/deposit/dataprep.pdf.

This guide provides an overview of data sharing considerations and procedures. Section 7 focuses on identifying and mitigating disclosure risk.

Inter-university Consortium for Political and Social Research. (2012). Guidelines for effective data management plans. Retrieved January 16, 2022 from https://www.icpsr.umich.edu/files/datamanagement/DataManagementPlans-All.pdf.

This guide describes elements to be included in data management plans, examples of each element, and links to additional resources.

Krenzke, T., Hubbell, K., Diallo, M., Gopinath, A., & Chen, S. (2014). Data coarsening and data swapping algorithms. Paper presented at the SAS Global Forum, Washington, DC. Retrieved January 16, 2022 from https://support.sas.com/resources/papers/proceedings14/1603-2014.pdf.

This paper demonstrates two proprietary SAS macros for coarsening variables and data swapping, illustrating concretely how coarsening and swapping works and some of the challenges in using the methods.

Schatschneider, C., Edwards, A., & Shero, J. (2021, July 30). De-identification guide. Figshare. Preprint. Retrieved January 16, 2022 from https://doi.org/10.6084/m9.figshare.13228664.

In addition to providing an overview of approaches for mitigating disclosure risk, this guide provides R, SAS, and SPSS code for examining variables for small cell sizes and recoding variables to coarsen them.

Shero, J. & Hart, S. (2020). Working with your IRB: Obtaining consent for open data sharing through consent forms and data use agreements. Figshare. Preprint. Retrieved January 16, 2022 from https://doi.org/10.6084/m9.figshare.13215305.v1.

This brief guide provides suggestions for language to include in participant consent forms and data use agreements that allows for data sharing.

Statistical software program

Templ, M., Kowarik, A., & Meindl, B. (2015). Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software.* Retrieved January 16, 2022 from https://www.jstatsoft.org/article/view/v067i04.

The developers of this package for the statistical program R describe its purpose as evaluating and anonymizing confidential micro-data sets. The package allows users to apply different methods of disclosure risk mitigation to their data, such as recoding, swapping, perturbing, and micro-aggregation. After implementing a method or combination of methods, users can assess both disclosure risk and information loss.

Disclosure risk management for qualitative text data

Inter-university Consortium for Political and Social Research. (n.d.) Guide to social science data preparation and archiving: Best practice throughout the data life cycle (6th edition). Retrieved January 16, 2022 from https://www.icpsr.umich.edu/files/deposit/dataprep.pdf.

This guide provides very basic guidance for sharing qualitative data (see pages 30 - 32 of the guide).

Social Science Research Council and Qualitative Data Repository. Managing qualitative social science data: An interactive online course. Retrieved January 16, 2022 from https://managing-qualitative-data.org/.

Module 3 of this on-demand online course addresses the sharing of qualitative data. A section of that module, Sharing Human Participant Data, covers disclosure risk issues for qualitative data.

Section II: Data documentation and organization

A guiding principle of data sharing is to provide something of value for science and, ultimately, for improving education. Clear, logical organization and careful documentation add considerable value to a dataset, while haphazard or incomplete attention to organization and documentation can render a dataset almost useless. This section provides suggested steps and resources for organizing study data files and developing a description of the data files and other supporting documentation. The first part provides suggestions and resources for how the original study team can organize and standardize its data management processes, ideally at the start of the study. The second part describes and illustrates the content and organization of a dataset ready to be shared with other investigators.

Organizing data management processes and documentation

At the beginning of a study, especially a multi-year project, sharing data might seem a long way off. In a sense, of course, that is correct, as data might not be ready to share until the main research findings are published—perhaps even after research funding has ended. However, the outset of a study is the ideal time for the study team to develop a plan for how it will manage its data: what the routines will be for organizing data files and code, how the team will regularly record important information about data collection and analytic choices, and which team member(s) will be responsible for overseeing this work.

The beginning of a study is also an ideal time for the principal investigator (PI) to express their commitment to sharing data and to set the expectation that team members will work carefully to organize and document data so that something of value can be shared with other investigators. PIs can also emphasize the more immediate benefits the team of clear, agreed-upon data management processes and a well-documented history of project decisions. These benefits include reducing duplication, inefficiency, and confusion through authoritative versions of data and code and easy access to previous versions of code. A written record of project decisions enables the team to retain information that is all too easily forgotten or lost when members move on to new roles or institutions.

Developing data management and documentation routines for a research team might take some investment of time and resources at the outset, but the benefits of having standard, organized procedures can extend beyond one study. Once established, these routines can be applied to multiple projects, with current team members socializing new recruits into how data and code are managed on the team.

Data management processes

Education research teams can benefit from applying best practices in data management developed by computer scientists and database experts, along with some specialized tools. As the numbers of sites, data sources, and team members grow, these systems become more important for maintaining consistency, enabling reproducibility of results, and avoiding mistakes such as overwriting original data sets. Practices address topics such as automating

the steps involved in creating and analyzing data, directory and file organization, version control of data and code, and creating self-documenting code.

Researchers with access to experienced data managers (for example, as a result of being affiliated with an academic center) are in a fortunate position. For researchers without such a resource, a first step to improve their data management strategy might be to consult a guide written for social scientists, such as Gentzkow and Shapiro's *Code and Data for the Social Sciences: A Practitioner's Guide* (see annotated bibliography at the end of this section for more information and link). University-based researchers might seek expertise in data management on campus, including through adding a researcher or graduate student in computer science, data science, or another related discipline to the research team. Researchers who are developing proposals for funding for new studies should consider including resources in the proposed budget to support data management and documentation, described next.

Documenting study decisions on an ongoing basis

A second step to take at the beginning of a study is to develop a plan for documenting study decisions. One strategy is to maintain a methods and measures (M&M) document: a single place to record information about the design, measures, and implementation of the study, including any anomalies that might be important to record. Box 7 describes a template for recording information about the study as it proceeds.

While the original research team is the primary audience for the M&M document, information from this document can be used to develop documentation for data sharing. Complete documentation can reduce the likelihood that the research team will receive calls and emails with questions from other investigators who are trying to use the data, which could potentially come months or years down the road.

Box 7: Methods and measures template

Link to Appendix A: Methods and measures (M&M) template

This template is closely modeled on one designed by Keith Smolkowski of the Oregon Research Institute. Used with permission.

The M&M template allows for narrative descriptions of the study design, participants, interventions, procedures, recruitment, consent processes, sample selection and maintenance, measures, and data analysis. These topics can be broken down to a more granular level, based on a study's needs and complexity.

For causal inference studies, some M&M sections can be initially populated with information from the study's pre-registration (for example, in registries such as the Registry of Efficacy and Effectiveness Studies and Open Science Framework Registries). The M&M document can also be used to record information needed to update a study registration if research questions or analytic plans change during the course of the study.

We recommend designating one research team member to update the file or direct other team members to update as needed.

If a research team has reached the middle of a study without an organized record of this information, all is not lost. The next step is to gather the team members, past and present, to recollect as best they can what took place and to share relevant documents. The team can use this information to populate a methods and measures document as described in Box 7.

Developing a data file structure and documentation for the shared dataset

Research teams who have followed good data management and documentation practices throughout the study will be well-prepared to assemble a complete dataset for sharing with other investigators. Preparing the dataset involves creating data files and documentation, including: (1) developing a data file structure; (2) reviewing programming code for disclosure risk and organization; (3) preparing a description of data files; and (4) providing documentation and/or code when data cannot be shared.

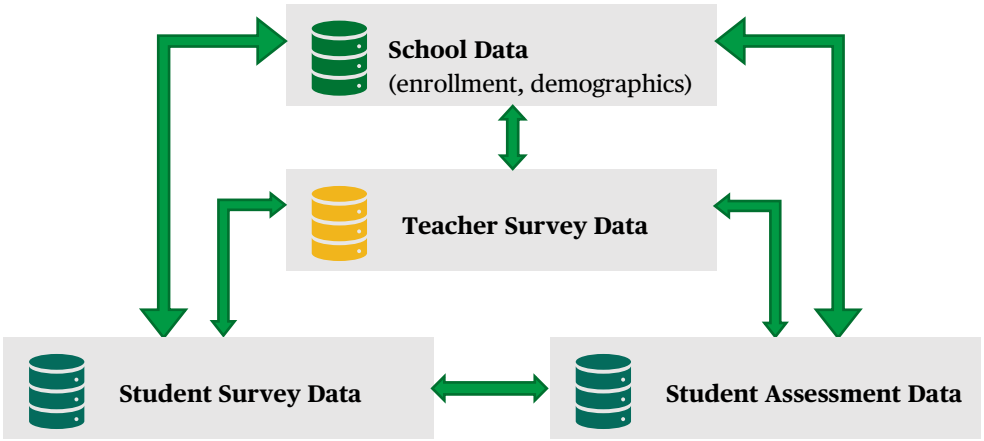
Step 1: Develop a data file structure

A research team that has carefully developed its own data file structure might have relatively little work to do to develop the file structure for the shared dataset. Datasets are typically comprised of multiple data files rather than a single large data file containing all variables and observations. Providing the data in multiple files makes it easier for the researcher to select only the variables needed for a particular analysis and minimizes computer processing problems associated with extremely large files. Each study should customize its file structure–that is, the number of files and their contents, as well as the linkages among them–based on the type of data and what will make it easiest for other researchers to use the data.

Education studies often collect data from people and units that are inter-related, such as students, teachers, principals, and school districts. Furthermore, data may be collected from these groups at several points in time. For a less complex study, such as one that collects data at one point in time and has just a few instruments, a file structure might be designed as in Figure 1. This file structure provides separate files for student, teacher, and school data, and includes two student files for the two types of student data collections (a survey and an assessment). With each file containing the ID of its own unit and all higher-level units, each file can be linked with every other file.

Data files should be shared in a file format, such as CSV or text, that is readable by a range of statistical packages. If the repository allows, researchers can also share files in the formats of their preferred statistical software.

Figure 1: A simple file structure with many ways to link files



Any file can be linked to any other file. Student files include student ID, teacher ID, and school ID. The teacher file includes teacher ID and school ID.

For studies with more complex data collections, such as longitudinal studies with multiple instruments, researchers should consider whether the benefits of creating separate files for each instrument and data collection point outweigh the drawbacks. We recommend that researchers developing file structures for complex studies seek to take the perspective of a new user of the dataset. For example, for a new user, a file structure with many separate data files might reduce potential confusion about which variables relate to a given round of data collection and which individuals provided data for the round. A disadvantage is the extra work of merging many files, especially if the content of the separate files could be just as easy to use if combined.

Regardless of whether the study data set is simple or complex, researchers should take care that variables follow a common, logical naming convention and have distinct names for any data elements that are included in more than one data collection. For example, if a study uses the same survey question at two points in time, the variable name for each administration of the question will need to be distinct.

Step 2: Review programming code for organization, clarity, completeness, and disclosure risk

Including programming code in the dataset can make the dataset more useful to other investigators and thus increase its value for the development of knowledge. For this reason, we recommend that researchers include statistical code that will help other investigators understand how the researcher created the analytic data files (including how data files were merged, variables constructed, and sample inclusion rules implemented). Code for executing analyses reported in scholarly publications can also be helpful to other investigators,

especially when the code includes annotations indicating where the analyses are reported in the publication (for example, a particular table or paragraph).

As with data files, code is more useful when it is organized into a set of programs that have clear and distinct purposes, and when each program is structured and documented in a way that makes it easy for another investigator to understand (see Box 10 for link to Appendix B.1, which contains a sample programming code structure).

Researchers should provide the code in whichever statistical program they use for their own analysis. If resources allow, researchers could consider also providing a document with pseudocode—that is, code specifications that describe the steps taken in plain language rather than in a specific statistical programming language. Pseudocode makes it easier for researchers who use other statistical programs to understand the steps taken to construct the file or conduct the analysis.

As with data files, researchers should carefully review the code they intend to share to ensure that it does not contain notes or other information that would directly identify individuals or clusters such as schools or districts. Particular care should be taken to ensure that the code used to clean and combine raw files does not include commands or explanatory notes that use participants' names or other direct identifiers.

Step 3: Prepare a Description of Data Files document

The Description of Data Files is the most essential piece of documentation in the dataset. It provides information about the study and explains the structure and contents of the data files (and code, if provided). At a minimum, this document should include an overview of the study design; description of data sources; sample sizes, response rates, and sampling procedures; measures and instruments used; variables; and file structure and programming code structure, including how files can be linked.

The Description of Data Files should also include three types of appendixes for each data file:

- **Codebook.** The codebook should include variable names, a brief description of the variable, the type of variable (such as numeric or text), and possible responses.
- **Annotated data collection instrument(s).** The instruments should be annotated with the variable name associated with each item. For instruments that cannot be shared (such as proprietary assessments), researchers should provide a brief description of the instrument, including the exact name and version; contact information for the instrument developer; and any other information, such as website links, that could be helpful for another researcher.
- **Summary statistics document.** For binary or categorical variables in the data file, this document should provide frequencies for various responses, as well as the number of study participants who did not respond to each item and, if available, the reason for nonresponse (for example, the participant did not know the answer). For continuous

variables, the document should report the number of study participants with nonmissing data, the mean value of the variable, and the minimum and maximum values.

Appendixes B.2–B.4 include a detailed sample outline for a Description of Data Files, examples of the three types of appendixes, and illustrative excerpts from a completed description of data files document for an impact study in education (Box 8).

Box 8: Resources for documenting and organizing data files and code

- Appendix B.1: Sample programming code file structure and contents for a shared dataset
- Appendix B.2: Sample outline for a Description of Data Files and instructions for each section
- Appendix B.3: Sample codebook, annotated data collection instrument, and summary statistics document
- Appendix B.4: Excerpts from a completed Description of Data Files for an impact study in education

Step 4: Provide documentation and/or code when data cannot be shared

Most researchers should at a minimum be able to share some of their study data using a restricted-use approach. However, when neither unrestricted nor restricted access data are possible (such as when an education agency will not agree to any form of data sharing), or when the data that can be provided is so minimal that it is unlikely to be useful to another investigator, researchers should consider making available documentation and/or code that can help another investigator reproduce or build on the analysis. For example, for studies that rely on education agency data, providing a list of variables requested from the agency and annotated code can give another investigator a head-start in understanding the variables and analytic choices of the original study. The code should be accompanied by information about how to obtain data from the education agency (for example, which office to contact, what forms to complete, and what the MOU will require). This approach supports replication and extension of research by allowing another investigator to request the same data from the agency and conduct analyses using the original researcher's code as a starting place.

Additional resources about data organization and documentation

Guides

Gentzkow, M. & Shapiro, J. (2014). Code and data for the social sciences: A practitioner's guide. University of Chicago mimeo. Retrieved January 16, 2022 from https://web.stanford.edu/~gentzkow/research/CodeAndData.pdf.

This is an extremely practical "how-to" guide for social scientists whose work involves coding but who are not formally trained in computer science. The guide covers topics like using directories to organize code files, version control, and self-documenting code. While not specifically geared toward preparing data for sharing, the guide can help researchers improve

their everyday programming, which in turn can increase the quality of code included in a shared dataset.

Inter-university Consortium on Social and Political Research. (n.d.). Guide to social science data preparation and archiving: Best practice throughout the data life cycle. 6th edition. Retrieved January 16, 2022 from https://www.icpsr.umich.edu/files/deposit/dataprep.pdf.

Section 5 of this guide, "Data Collection and File Creation," provides detail on organizing data files, naming variables, and handling data that are missing, imputed, geospatial, or qualitative.

Reynolds, T. & Schatschneider, C. (2020). The basics of data management. figshare. Preprint. Retrieved January 16, 2022 from https://doi.org/10.6084/m9.figshare.13215350.v1.

This paper provides practical ideas for organizing data collection and data entry, planning data files, and maintaining procedures for version control.

PowerPoint

Logan, Jessica. (2019). Data management and data management plans. Figshare. Presentation. Retrieved January 16, 2022 from https://doi.org/10.6084/m9.figshare.7890827.v1.

This PowerPoint provides practical strategies and examples for organizing and documenting data with data sharing in mind.

Video presentations and podcasts

Logan, Jessica, and Sara Hart, speakers. (2020, July 8). The what, why, how, and hesitations of data sharing. *Within & Between: A Developmental Science Podcast*. Retrieved January 16, 2022 from https://anchor.fm/withinandbetween/episodes/Episode-4-The-what--why--how--and-hesitations-of-data-sharing-efu17q/a-a2k3to9.

In this podcast, Jessica Logan and Sara Hart describe the motivations for sharing data, the importance of good documentation, and what the documentation should include.

Useful techniques for sharing data effectively. Recorded panel presentation at the 2020 IES Principal Investigators Conference, Washington, D.C. Retrieved January 16, 2022 from: https://ies.ed.gov/pimeeting/2020/SessionMaterials.aspx.

IES grantees Margaret Burchinal (University of North Carolina) and Jessica Logan (Ohio State University) describe data organization strategies and quality assurance practices and, along with panel member Terry Sabol (Northwestern University), answer audience questions about data sharing.

Organizing and documenting qualitative data

Social Science Research Council and Qualitative Data Repository. Managing qualitative social science data: An interactive online course. Retrieved January 16, 2022 from https://managing-qualitative-data.org/.

Modules 1 and 2 of this on-demand online course address management and documentation of qualitative data.

Section III: Depositing the dataset

There are several main types of approaches that researchers can take to share their study data. The main options, which can be used alone or in combination, include:

- Depositing data in a data repository–a centralized web-based or physical location for sharing data, also known as a data archive;
- Providing data using a personal website, a university data archive, or corporate website. Public-use files can be placed on the website for anyone to download, or the website can provide information about the content of the dataset and directions for requesting access; and
- Developing an agreement with the agency or post-secondary institution that originally provided data to host the dataset and make it available for qualified investigators to use.

This section summarizes issues for researchers to consider when investigating and weighing these general alternatives for where to share their data, as well as choosing among specific options. Each approach has benefits and drawbacks, and we provide examples of education researchers who made different choices.

Data repositories

We recommend that researchers first explore the option of placing their study data in a data repository. Repositories offer several important advantages over other approaches. Sharing data through a data repository, rather than through a personal website or an education agency, may increase the likelihood that another researcher who is interested in the topic will discover the data. For this reason, repositories that are commonly used by researchers in the field or that aim to build a community around the repository may be especially good options. Further, some data repositories offer sophisticated options for accessing data securely and support for preparing data to be shared.

However, data repositories differ in the options, services, and supports they offer for data sharing. Researchers should consider whether repositories of interest are suitable based on dimensions such as:

- **The likelihood that other investigators in the field will learn about the dataset.** To contribute to the development of knowledge, shared datasets need to be discovered by others with relevant interests. For this reason, researchers should consider whether colleagues in their field are likely to learn about the dataset if they share it through a particular repository. Repositories that are unfamiliar to most members of a scholarly community that would be interested in the dataset might not be the best choice to facilitate discovery.
- **The cost to store or access the data.** When planning a research project, researchers should consider the cost to their project of sharing study data through a data repository so that the cost can be factored into the study's budget. The cost to other researchers who

seek to access the data should also be taken into account, as the cost could limit the potential audience.

- **The types of data and formats accepted.** Some data repositories only allow for storage and sharing of quantitative or qualitative data. Studies that include qualitative data, images, or audio or video recordings will require a data repository that allows for these types of data.
- **Options for accessing data.** It is common for data repositories to allow users to download data sets from the repository website, but some repositories offer more sophisticated options. For example, virtual data enclaves add another layer of data security by allowing researchers to analyze data remotely, without downloading data to their computer. For very sensitive data, on-site data enclaves allow researchers to use the data at a secure physical site monitored by the repository.
- **The size limits of data upload.** The maximum file size for data and documentation varies across repositories. If files are exceptionally large (over 5 GB), researchers should be aware of the limits that might impede upload capabilities. The total space available for a deposit may differ from the space allowed per file. For example, a repository might allow 5 GB of storage but a maximum individual file size of only 100 MB.
- **Assignment of a Direct Object Identifier (DOI) and citation.** A DOI is a persistent identifier for an electronic object. When a repository assigns a DOI to the shared dataset, the researcher can track how the dataset contributes to the development of knowledge through other researchers' published papers and citations.
- **The length of deposit.** Recipients of federal research grants are expected to provide their data for 10 years, at a minimum. If a data repository does not charge for either depositing or accessing the data, consider whether the data repository has other funding sources that will enable it to store data securely and make the data available for at least a decade.
- **Depositors' rights**. Understanding the rights and responsibilities of the data repository, as well as those of the researcher, is critical. Researchers often maintain the rights to their data when submitting them to a data repository. Researchers should also understand whether the repository gives them the right to remove their data or to replace it with a dataset that has been added to or amended.
- **Data cleaning and curating options.** For a fee, data repositories may offer disclosure risk reviews and a variety of data cleaning and organizing services. If a study budget allows, these services can provide additional assurance to the researcher that disclosure risk is being managed appropriately and that the researcher is providing a well-organized, high-quality dataset. Repositories allowing restricted-use files should provide evidence that any repository staff handling the data are trained in handling files of this type.
- **Independent certification of trustworthiness.** When sharing their study data through a repository, researchers should be confident that the repository has high standards for data quality, can be expected to provide the data over time (that is, it is sustainable), and has appropriate procedures to protect access to the data. Some repositories are certified

as a Trustworthy Data Repository by CoreTrustSeal, an international organization that promotes sustainable and trustworthy data infrastructures.

Data repositories of special interest to education researchers

The following data repositories have a special focus on data from education or social science studies or are prominent repositories that encompass data from a wide range of disciplines. The repositories are presented in alphabetical order. Table 2 provides key information about these repositories.

- **Harvard Dataverse** (dataverse.harvard.edu). This platform allows researchers from any discipline, institution, or organization to share their data or to set up collections of related datasets (a "dataverse"). There is no cost to the researcher to add datasets, which can be either unrestricted or restricted. Researchers can set the terms of use for restricted access files, including an option of requiring that interested potential users obtain permission from the researcher to access the data. Note that some university dataverses use the Harvard Dataverse platform.
- **Inter-university Consortium for Political and Social Research (ICPSR)** (www.icpsr.umich.edu). ICPSR has a 50-year history of serving as a data archive and is certified as a Trustworthy Data Repository by CoreTrustSeal. ICPSR provides a variety of options for social science researchers who want to deposit data and for investigators who want to use the data.
  - **Data curated by ICPSR.** ICPSR will, at a minimum, review study data for disclosure risk and make changes as needed to preserve confidentiality. It can also improve documentation and usability of the dataset. However, there are costs and access to consider. If the researcher pays for ICPSR's curating, the dataset will be freely available; if the researcher does not pay for curating, the dataset will be available only to individuals from institutions that are ICPSR members.
  - **OpenICPSR** (https://www.openicpsr.org/openicpsr/). This ICPSR site allows researchers to share their datasets at no cost to the researcher, though data users may be charged for access to the data. ICPSR does not check data uploaded to OpenICPSR for errors or confidentiality issues, so it is the responsibility of the researcher to ensure that the dataset is well-organized, properly documented, and maintains participant confidentiality. OpenICPSR accepts both restricted and unrestricted data.
- **LDBase** (ldbase.org) LDBase is a new data repository hosted by Florida State University with funding from the National Institutes of Health. LDBase will accept data from education studies involving students across a range of abilities. It does not expect to charge a fee for depositing data.
- **Open Science Framework (OSF)** (https://osf.io/) OSF is a multi-faceted research platform that supports scholarly collaboration and sharing of research materials. Researchers can decide what material from their studies, including data files, to share publicly. OSF does not offer services to prepare datasets for release.

Table 2: At-a-glance summary of data repositories

	Harvard DataVerse	Inter-university Consortium for Political and Social Research (ICPSR)		LDBase	Open Science Framework
		Data curated by ICPSR	OpenICPSR		
Costs					
Free to deposit data	✓	✓	✓	✓	✓
Free to access data	✓	Yes, if researcher pays ICPSR to curate data; otherwise, only affiliates of member institutions ($550 per study for non-members)	ICPSR may charge a fee to access data	✓	✓
Types of Access					
Accepts restricted-use files	✓	✓	✓	✓	X
Offers virtual data enclave	X	✓	✓	X	X
Offers on-site data enclave	X	✓	✓	X	X
File/data types					
File size limit	2.5GB	None specified	2GB	None specified	5GB
Dataset size limit	1TB	None specified	2GB	None specified	No limit
Other features					
Assigns DOI	✓	✓	✓	✓	✓
Researcher can update their data	✓		✓	Not specified	✓
Researcher can remove their data	✓		✓	Not specified	✓
Offers data curation services	✓	✓	✓	X	X
Certified as Trustworthy Data Repository by CoreTrustSeal	X	✓	✓	X	X

The information from this table was compiled as of January 2022. Researchers should consult each data repository's website to obtain the most updated features of each repository.

Alternatives to data repositories

The variety of data repositories available to education researchers, including several no-cost options, means that most education researchers should be able to find a repository that meets their needs and their study budget.

Nevertheless, researchers do encounter situations where they cannot share their data through a repository. These situations call for thinking creatively about other ways to enable others to discover and access the data.

Education researchers whose studies include education agency administrative data may find that the agency is unwilling to allow its data to be shared, even in a restricted-use format. Some agencies are uncomfortable allowing a researcher they do not know or trust to access their data. They may fear not only disclosure risk but also new, unexpected analyses to which they would be unprepared to respond.

In some cases, conversations with the agency about disclosure risk mitigation strategies will be sufficient to increase the agency's comfort level with sharing the study data through a repository. If their concerns remain, another option is to provide the dataset to the agency, which can grant access for its use through its regular research application procedures (see Box 9 for an example). The dataset should be prepared as it would be for a data repository, including mitigating disclosure risk and carefully documenting the data files.

Researchers who provide their study data to an agency for others to use will need to take steps to let others know that the dataset is available. This information can be shared on a personal website or added to notes in a manuscript. Minimally, a description of the study data on a website should include the project overview from the Description of Data Files and information about how to contact the agency to request access to the data.

Box 9: Example of depositing study data with a school district

A study dataset can be given to an agency that provided the original data, and the agency can pledge to maintain the data for future investigators to use. Then, on a website or through another venue, investigators can share information about how to access the dataset, including whom to contact and which files to request, as well as details about the variables they used and the analytic steps they took. Investigators could also share their code as part of the dataset kept at the agency or as a file that can be downloaded or requested.

For example, Regional Educational Laboratory Mid-Atlantic provides a document detailing the data it used for its impact study of home visiting in the District of Columbia. This document is available on a study landing page that also includes the study report, study summary, and technical appendixes. The document provides information about how to request the analytic files from the District of Columbia Public Schools, a description of the relevant files and variables, and a brief summary of how the data were cleaned, merged, and analyzed.

When to deposit data

Considering the effort and resources involved in organizing data files and preparing data documentation, researchers should consider how many datasets to release from a study and when to release the dataset(s). A basic standard for the timing of data sharing is to make the data available when the study's main findings are accepted for publication in a peer-reviewed scholarly journal or published in another way, such as by a government agency after peer review. For causal studies, a reasonable definition for a main finding is one that answers a confirmatory research question as described in the study's entry in a pre-registration database (such as the Registry of Efficacy and Effectiveness Studies or the Open Science Framework Registries). IES-funded researchers should keep in mind that IES expects study data to be made available concurrent with the relevant publication's acceptance to a peer-reviewed scholarly journal.

Data repositories typically allow updated versions of a dataset. When researchers publish main findings in separate manuscripts, they will need to decide whether to use the option of updating a previously-released dataset or make a separate dataset available that corresponds to the new publication. Each research team needs to develop an approach that balances efficiency and the utility of the datasets for other investigators. For example, if a later study publication adds only a small number of new variables, it may be easiest for the researcher and more useful to other investigators to issue an updated dataset. Care will need to be taken not to introduce inconsistencies between versions of the data files that make it difficult to reproduce findings from various manuscripts.

If study datasets are released separately, researchers will need to consider whether to include unique identifiers that allow the datasets to be merged. Regardless of whether the same identifiers are included to allow merging, special care should be taken to ensure that new data files do not inadvertently create disclosure risks.

Additional resource about depositing the dataset

Reference document

Inter-university Consortium for Political and Social Research (ICPSR). (2020, September). ICPSR curation levels. Retrieved January 16, 2022 from https://www.icpsr.umich.edu/files/datamanagement/icpsr-curation-levels.pdf.

This document describes activities that ICPSR conducts for three levels of data curation and the approximate number of weeks from the start of curation to data release.

Section IV: Checklist of key steps for data sharing

There is a lot to keep in mind when preparing a study dataset for sharing. This checklist summarizes key issues for researchers to consider and steps to take, as discussed in this guide. The checklist is organized according to the sections and sub-sections of the guide.

Section I: Disclosure risk management

Authorization to share data

- ☐ Does the researcher have authorization from data owners, such as education agencies or postsecondary institutions, to share study data in any form?
  - ☐ Have any data owners provided explicit permission to include their data in a shared data set?
  - ☐ Does the planned data sharing approach abide by any limitations from data owners on how the data may be shared?
- ☐ Do the study's participant consent forms allow for data sharing?
- ☐ Has the Institutional Review Board assessed whether data from the study may be shared?
- ☐ (For new studies) Has the researcher communicated transparently with agency or institution partners about plans for data sharing?

Assessing Disclosure Risk

- ☐ Have the data files been reviewed carefully for variables that could directly or indirectly identify individuals, including by combining the data with information from public sources?
- ☐ Have combinations of indirect identifiers with fewer than three cases per cell been identified?
- ☐ Have variables of a sensitive nature been flagged?
- ☐ If raw and analytic data files will be shared, have both types of files been reviewed for disclosure risk, including risks from merging the files?

Mitigating Disclosure Risk

- ☐ Have direct identifiers been removed from the data files?
- ☐ For indirect identifiers or variables of a sensitive nature: can these be removed or modified to mask identities?

- ☐ Has the researcher assessed whether these changes significantly affect the utility of the dataset for replication and/or extension?
- ☐ Can any of the data be shared in a public-use format while retaining their utility for reproducing original study results and/or extending the analysis?
  - ☐ If some, but not all, of the data can be shared in a public-use format, can a combination of restricted-use and public-use files be used?
- ☐ If the data cannot be shared in a public-use format and retain their utility, can they be shared in a restricted-use format?
- ☐ Will a restricted-use file also need indirect identifiers or variables of a sensitive nature to be modified for additional disclosure risk mitigation?
- ☐ Taking into account the disclosure risk and potential mitigation challenges of the data, should the team consult an expert in disclosure risk mitigation?

Section II: Data documentation and organization

Data Management and Documentation (beginning of a study)

- ☐ Has the principal investigator clarified team members' roles and responsibilities for ongoing documentation of study methods and measures?
  - ☐ Is there a point-person who is responsible for ensuring that documentation is complete and up-to-date while the study is ongoing?
- ☐ Does the study team include an experienced data manager? If not, has the team consulted an expert in database organization and management for help with developing a file structure and data management procedures?

Data File Structure

- ☐ Has a logical data file structure been developed that would facilitate another investigator's use of the data?
- ☐ Does the file structure take into account the particular characteristics of the study, such as the number of data sources and data collections?
- ☐ Are there multiple ways to link files to each other, appropriate to the file structure?
- ☐ Are variables named according to a logical, common naming convention, and are data elements that are part of more than one data collection given distinct names?

Developing a Programming Code Structure

- ☐ Has programming code been identified that could be useful for a new user and included in the dataset?
- ☐ Has a structure been developed for the programming code?
- ☐ Is the programming code well-organized and annotated to be informative to a new user?

Preparing a Description of Data Files

- ☐ Has a Description of Data Files document been developed to orient a new user to the study and the data being shared?
- ☐ Does the Description of Data Files document include information about:
  - ☐ The study design
  - ☐ Data sources
  - ☐ Sample sizes, response rates, and sampling procedures
  - ☐ Measures and instruments used
  - ☐ Names and descriptions of variables
  - ☐ Data file structure and programming code structure, including how files can be linked
- ☐ For each data file, does the Description of Data Files include an appendix with a codebook, annotated data collection instruments, and frequencies and summary statistics?

Other Issues

- ☐ If no data can be shared, have options been explored for making available programming code and documentation that another investigator can use if they can obtain access to the data?

Section III: Depositing the Dataset

Exploring Options

- ☐ Have options been considered for when the data will be shared, including how many separate datasets will be released from the study?
- ☐ If data owners will not allow data to be shared in a data repository, are they willing to host it and make it available for other researchers to apply to use?

- ☐ If considering sharing data through a personal or university website, have the following been assessed?
  - ☐ A strategy for informing researchers that the data are available
  - ☐ A plan for how researchers will apply to access the data and how they will be vetted
  - ☐ Agreements that the potential users will need to sign
  - ☐ A plan for how the data will be securely provided to approved users
  - ☐ The time required to manage requests for data and agreements

Choosing a Data Repository

- ☐ Have potential data repositories been assessed for:
  - ☐ Likelihood that researchers who might be interested in the data would discover it, including whether a DOI will be assigned
  - ☐ Cost to deposit and/or maintain, and cost to access
  - ☐ Likelihood of being able to maintain the data for at least 10 years
  - ☐ Limits on types of data or file sizes
  - ☐ Support available for preparing datasets for sharing
  - ☐ Independent certification of trustworthiness

References

Clark, M., Chiang, H., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). The effectiveness of secondary math teachers from Teach For America and the Teaching Fellows Programs (NCEE 2013-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved January 16, 2022 from https://ies.ed.gov/ncee/pubs/20134015/pdf/20134015.pdf.

Holdren, J. (2013). Increasing access to the results of federally funded scientific research. U.S. Office of Science and Technology Policy. Retrieved January 16, 2022 from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

National Center for Education Statistics. (2010). Basic concepts and definitions for privacy and confidentiality in student education records (SLDS Technical Brief 1). Retrieved April 29, 2021 from https://nces.ed.gov/pubs2011/2011601.pdf.

National Center for Education Statistics. (2002). Statistical standards. Retrieved April 29, 2021 from https://nces.ed.gov/statprog/2002/std4_2.asp.

Appendix A: Methods and measures template

The Methods and Measures (M&M) template provides a format for a living document that researchers can use to record information about their projects. The M&M document contains detailed descriptions of procedures, measures, methods, and sample, as well as any changes that occur during the lifetime of a project. It can assist with data sharing by tracking the details that can easily be forgotten as time passes and research staff come and go.

Return to Section II: Data Documentation and Organization

The M&M document can be initially populated with information provided in the study's pre-registration. To keep the M&M document organized and up-to-date, the document should be managed by a designated investigator or project manager who directs staff to update sections or collects and adds text during the course of the project.

This M&M template is designed to be used for group design causal inference studies, especially those that involve testing an intervention in an education context. The document can also help the team to record data that the What Works Clearinghouse (WWC) needs to conduct a study review. The WWC provides a Reporting Guide for Study Authors summarizing the descriptive information and study data used to assess studies against standards. Tables from the guide are reproduced at the end of this M&M template.

Researchers who are conducting studies other than group design causal inference studies may be able to adapt sections of this template to suit their studies.

This M&M template is adapted with permission from a template developed by Keith Smolkowski of the University of Oregon.

Project Description and Research Design Overview

Provide an overview of the project. The overview could include relevant parts of the abstract from a grant application or study pre-registration, such as the key research questions, motivation, and a brief description of the study design. Detail about the study design, such as the original plan for participants and methods, should be placed in their respective sections below.

Participants

Describe the study participants, first as planned in the original project description, and then any changes that occur during the course of the project (such as units being added or dropped). Organize this description by relevant units, as illustrated below (headings are for a K-12 study and should be modified to be appropriate for the specific study).

This section can include tables with numbers of units and participants and key information about them, such as demographic characteristics. Some data from this section may be needed by the WWC to conduct a study review (see tables at the end of this M&M template).

A. Districts

B. Schools

C. Teachers and School Staff

D. Students

Intervention(s)

Describe the intervention(s) with a focus on implementation and activities, first as planned in the original project description, and then any changes that occur during the course of the project. If units adapted the intervention, identify these adaptations and the units that implemented them.

Procedures

In this section, describe procedures used to implement the study, first as planned in the original project description, and then any changes that occur during the course of the project. If a section of procedures is not relevant to a study, the section can be deleted.

Recruitment

Describe how participants were recruited to participate in the study. This description can include recruitment procedures for different units (for example, districts, schools, and teachers).

Permission and Consent

Describe the procedures used to obtain permission to conduct research (for example, submitting a research application to districts or schools) and obtaining consent from individual participants (such as staff members, parents, or students).

Sample Selection

Explain how units and/or individuals were selected to participate, including the criteria for selection and any screening to identify whether individuals met the criteria.

Randomization

Explain procedures used to assign units to intervention(s) and control groups.

Sample Maintenance

Describe any procedures for retaining participants in the study over time, such as efforts to identify participants' contact information and build motivation for continuing to participate.

Assessments and Observations

Briefly describe assessments used and the procedures for administering these assessments. Measures will be described in more detail in the "Measures" section below.

Observations

Describe any observations conducted, including any rubrics used and the procedures for conducting these observations.

Measures

Describe the measures, first as planned in the original project description, and then any changes that occur during the course of the project. Include the reliability of the measures, if applicable and known. For constructed variables, indicate how they were defined. This can also be a place to record known difficulties with measures, such as school administrative data that is known to be inconsistently reported or errors on survey forms. In this section, you can also include the location (drive, file name) of code for creating the variables.

Measures can be organized by key units in the study, as illustrated below for a post-secondary study.

A. Schools

B. Transition programs

C. Students

Data Analysis

Describe the data analysis, first as planned in the original project description, and then any changes that occur during the course of the project. Some data from this section may be needed by the WWC to conduct a study review (see tables at the end of this M&M template).

Impact Analysis

This section includes a narrative about how the analytic sample was constructed, including how missing data was handled. The section can include links to programming code used to construct the analytic file and perform the analysis. Detailed information about outcomes, missing data, and baseline equivalence for the analytic sample can be summarized in tables at the end of this M&M template.

Additional analyses

Narrative descriptions of, and data from, exploratory analyses and cost analyses can be included in the M&M file as needed.

Summary of Information Needed for What Works Clearinghouse reviews

The tables in this section are reproduced from the What Works Clearinghouse Reporting Guide for Study Authors, which also includes more information about completing each table. Not all tables are relevant to every study. For ease of reference, we recommend creating separate sets of tables for each manuscript.

Table A.1: Information to include for each outcome measure, time point, and comparison

	Intervention Group			Comparison Group			Estimated Effect		
	Sample Size	Mean	SD	Sample Size	Mean	SD	Estimate	p-value	Effect Size
Outcome measure									
Pre-intervention measure									
Other key pre-intervention characteristic 1									
Other key pre-intervention characteristic 2									

Table A.2: Additional data to include for RCTs that assigned individuals to the intervention and comparison groups

Outcome Measure	Intervention group sample size at random assignment	Comparison group sample size at random assignment
Measure 1		
Measure 2		
Measure 3		

Table A.3: Additional sample sizes to include for all studies that assigned clusters to the intervention and comparison groups

	Intervention group			Comparison group		
		Individuals within clusters that remain in the analytic sample			Individuals within clusters that remain in the analytic sample	
	Number of clusters that remain in the analytic sample	Around the time pre-intervention data were collected	Around the time outcome data were collected	*Number of clusters that remain in the analytic sample*	Around the time pre-intervention data were collected	Around the time outcome data were collected
Outcome measure 1						
Outcome measure **2**						
Outcome measure **3**						

Table A.4: Additional sample sizes to include for RCTs that assigned clusters to the intervention and comparison groups

	Intervention group		Comparison group	
	Number of clusters randomly assigned to condition	At the earliest point in time after all joiners had entered clusters, total number of individuals in clusters that remain in the analytic sample	Number of clusters randomly assigned to condition	At the earliest point in time after all joiners had entered clusters, total number of individuals in clusters that remain in the analytic sample
Outcome measure 1				
Outcome measure 2				
Outcome measure 3				

Table A.5: Data to include for each pre-intervention measure for which any observations are missing or imputed in the analytic sample and each outcome measure for which any observations are imputed

	Intervention group			Comparison group		
Subsample of analytic sample	Number of observations	Mean of pre-intervention measure	Mean of outcome measure	Number of observations	Mean of pre-intervention measure	Mean of outcome measure
Units for which both outcome and pre-intervention measures are observed						
Units for which only outcome measure is observed		not applicable			not applicable	
Units for which only pre-intervention measure is observed			not applicable			not applicable
Correlation between the pre-intervention and outcome measures						

Appendix B.1: Sample programming code file structure and contents

When preparing a dataset for sharing, it is important to consider what will best help another researcher to understand the data and begin to use it with a modest amount of effort. Datasets that include programming code have additional value beyond those that include only data files and documentation. When the code files are structured in a clear and logical manner and employ some basic coding best practices, the dataset is even more useful to other researchers who are seeking to reproduce or extend the original analyses.

Return to resources for documenting and organizing data files and code

This appendix describes steps that researchers can take to organize their programming code files and improve the quality of the code. If used across the study life cycle, these practices can improve the original study team's data management and make it simple to include useful code in the shared dataset.

Organizing programming code files

Programming "modules." We recommend creating separate programming code files for major steps in data cleaning, preparation of analytic data files, and analysis. While it can be tempting to create a single programming file so that all of the code is in one place, separate code "modules" that have distinct purposes and are clearly labeled are preferable. If code needs to be modified, the programmer should be able to go right to the relevant programming file to make changes rather than trying to navigate through a single file that may be extremely long. Programs should be numbered to indicate the order in which they are to be executed.

Directories. Programs, as well as the associated raw files, output, and analytic files, should be organized into directories corresponding to stages of data preparation and analysis. This directory structure can be used to organize both programming code files and data files in the shared dataset. The original research team might have directories for cleaning data, creating an analytic file, analysis, and reporting (formatting tables for a manuscript). In a shared dataset, the team might decide to include just the directories related to creating an analytic file and analysis.

For example, a directory structure might be organized as in Box B.1.1. Illustrative examples of programming code files and data files are shown in each directory.

Box B.1.1: Organize programming code files and data files in directories

1. **./01_ Raw**
   a. **Student**
      i. Student_survey_wave1.csv
      ii. Student_test_wave1.csv
      iii. Student_survey_wave2.csv
      iv. Student_test_wave2.csv
   b. **Teacher and School**
      i. Teacher_survey_wave1.csv
      ii. Teacher_survey_wave2.csv
   c. **School**
      i. School.csv
2. **./02_ Analytic**
   a. **Programs**
      i. 01_merge.do
      ii. 02_constructs.do
      iii. 03_impute.do
   b. **Code specifications**
      i. Specifications.doc
   c. **Output**
      i. **Student**
         1. Student_analysis.csv
         2. Student_analysis.dta
         3. Student_imputed.csv
         4. Student_imputed.dta
      ii. **Teacher and school**
         1. Teacher_analysis.dta
         2. Teacher_analysis.csv
         3. School_analysis.dta
         4. School_analysis.csv
3. **./03_ Analysis and Reporting**
   a. **Programs**
      i. Student_analysis.do
      ii. Teacher_analysis.do

Raw data files are organized by sub-directories, customized to the study

Provide raw data files in a widely readable format like CSV or text

Programs are in the preferred statistical package of the original research team and numbered by the order in which they are executed

Specifications for constructed variables and sample inclusion rules, written in "plain English," not syntax

Analytic files provided in widely readable format (CSV, text); files in other statistical software formats optional

Programs with analytic code show how estimation approach was executed

Best practices for organizing and documenting code

Researchers should make an effort to share well-organized, visually uncluttered code. A few basic practices will go a long way toward improving the quality of the code in a shared dataset. These are:

Meaningful and clear documentation in the code

- Include a header box in each program (using comment syntax) that includes date created, purpose and description, inputs, and outputs.
- Provide meaningful and clear comments throughout the code.
  - For example: At the start of each new step, it is useful to insert a comment indicating what actions are being performed within that step. For long and/or complex programs, add comments to clearly denote major sections of code.
  - For example: When creating a complex variable or one that requires several lines of code, add a comment to describe the process used to create the variable.
  - Make sure to update comments if the specifications change.

Structuring code for readability and visual clarity

- Separate blocks of code with blank lines so that they are more readable.
- Indent code where appropriate to show the structure of the code. Indenting makes it easier for a new user to read and follow the logic of the code.

Appendix B.2: Sample outline for a description of data files document

One of the benefits of maintaining a thorough and up-to-date Methods and Measures document is that many parts of it can be used with very little revision to complete the Description of Data Files.

Return to resources for documenting and organizing data files and code

I. Introduction

In 2-3 paragraphs, this section should provide a brief overview of the study and describe the content and organization of the document. The Project Overview section of the Methods and Measures template can be a source of information for the study description. This section should also include a citation and link for published manuscript(s) based on the data. IES-funded researchers should link to the ERIC citation of the relevant manuscript(s).

II. Study Design

The section includes several subsections, each of which can be addressed in 1-2 short paragraphs. Subsections can be added as needed to describe important features of the study design.

A. Research questions

B. Experimental design (as applicable)

C. Sample and sampling design (as applicable)

D. Recruitment process (as applicable)

a. Table(s) of sample sizes should be included in this subsection.

III. Data Sources

In this section, describe each data source and/or each unit of analysis in a separate subsection. Examples of data sources include teacher surveys, school records, classroom observations. Examples of units of analysis include teachers, principals, or students.

- For each data source/unit of analysis:
  - Identify the unit of analysis if the section is organized by data source.
  - Describe types of information collected (for example, "information on educational attainment and classroom practices"). Do not list each individual data item in this section.
  - Describe how data were collected (for example, paper or web survey, administrative records, classroom observations). Note important aspects of surveys, such as whether student surveys were administered by classroom teachers, outside data collectors, etc.

  - State what time period the data cover.
  - State when data were collected. If there were multiple cohorts or more than two rounds of data collection, include a chart showing the data collection schedule.
  - Include summary data collection statistics (for example, sample sizes and response rates for each source, sample attrition information, as applicable). For experiments, statistics need not be broken out by treatment status, unless there are important differences by status.

IV. Data Files

This section provides considerable detail on the organization and content of the data files in the shared dataset. A potential set of subsections is as follows:

Data file preparation

- Describe any important aspects of data entry or data cleaning.
- Describe any variable or file naming conventions that would be helpful for users to know.

Detailed descriptions of files or related groups of files

- **Summary.** Provide a summary description of how data from multiple data sources were organized and combined into data file(s), including organization of sample information variables (IDs, treatment status, cohort flags, weights), raw (or source) data, constructed (or derived or composite) variables.
- **Sample information variables.** List and describe sample information variables, such as treatment or random assignment status.
- **Weights.** Describe how sampling and/or non-response weights were defined and how they should be used in analyses.
- **Constructed variables**
  - Provide a general description of methods used. Refer to an appendix for detailed information on each constructed variable.
  - It is not necessary to describe simple recodes (e.g., gender being recoded from F/M to 1/2) or to state that ID numbers were used in place of names.

Missing data

- Describe how missing data were handled, including imputation procedures, if applicable.
- This section does not need to describe variable suppression or masking procedures applied to protect confidentiality.

Data file details

- Descriptive information on each data file

  - This should include the unit of observation and number of records (which might be different from the sample sizes given in the Study Design or Data Sources sections).
- Instructions for combining data files
  - Key ID variables that link datasets
  - Sample code for merging the data files (optional)
  - List missing value codes that are used in the data files.
- List of files included in the dataset
  - Include actual file names, with directory structure (if applicable)
  - Include both data and documentation files (such as instruments, codebooks, and summary statistics)

V. Appendix

Three documents for each data file should be included in the appendix: a codebook, annotated data collection instruments (or a description of the instrument if it cannot be shared), and summary statistics. See Appendix B.3 for examples of these appendix materials.

Appendix B.3: Sample codebook, annotated data collection instrument, and summary statistics document

For each data file, the Description of Data Files document should include three appendixes: a codebook, annotated data collection instrument(s), and frequencies and summary statistics. This appendix provides illustrative examples of content for documents.

Return to resources for documenting and organizing data files and code

Codebook

The codebook should include variable names, a brief description of the variable, the type of variable (such as numeric or text), and possible responses. Table B.3.1 provides an example for a data file with school-level variables.

There are some useful variable naming conventions to note. The variable names begin with *sc_* to indicate that they are school-level variables. Variables that provide similar information for different school years indicate the year in the last four digits (for example, *sc_withdrew1718*). Variables containing data from the Common Core of Data, a publicly available dataset provided by the National Center for Education Statistics, begin with *sc_ccd* to indicate their provenance.

Table B.3.1: Codebook example for a school-level data file

Variable name	Description	Type	Values
school_id	School identification number	Numeric	All whole numbers between 1 and 100 (study generated)
sc_enrollment_year	School year began participating in the study	Numeric	1=2017/18 2=2018/19
sc_treatment	School was assigned to implement the study intervention	Numeric	0=No 1=Yes
sc_withdrew1718	School withdrew from the study during the 2017/18 school year	Numeric	0=No 1=Yes 9997=Not Applicable (began study participation in 2018/19)
sc_withdrew1819	School withdrew from the study during the 2018/19 school year	Numeric	0=No 1=Yes
sc_ccd_title1	School was Title I eligible during the year before study enrollment	Numeric	0=No 1=Yes 9998=Don't know
sc_ccd_frp	Percent of school's students eligible for free or reduced-price meals during the year before study enrollment	Numeric	0–100 9999=Refused

Table adapted from Agodini, R. (2015). Guide for constructing a restricted-use data file for NCEE studies. Unpublished manuscript.

Annotated data collection instrument

Researchers should annotate their data collection instruments by indicating how the variable names in the data file correspond to items on the instruments. Annotations can be made on a formatted copy of the instrument, as illustrated below. In this example variables that are not included in the data file (in this case, qualitative responses to questions 1 and 2, and teacher names provided in question 3) are clearly identified as not part of the dataset.

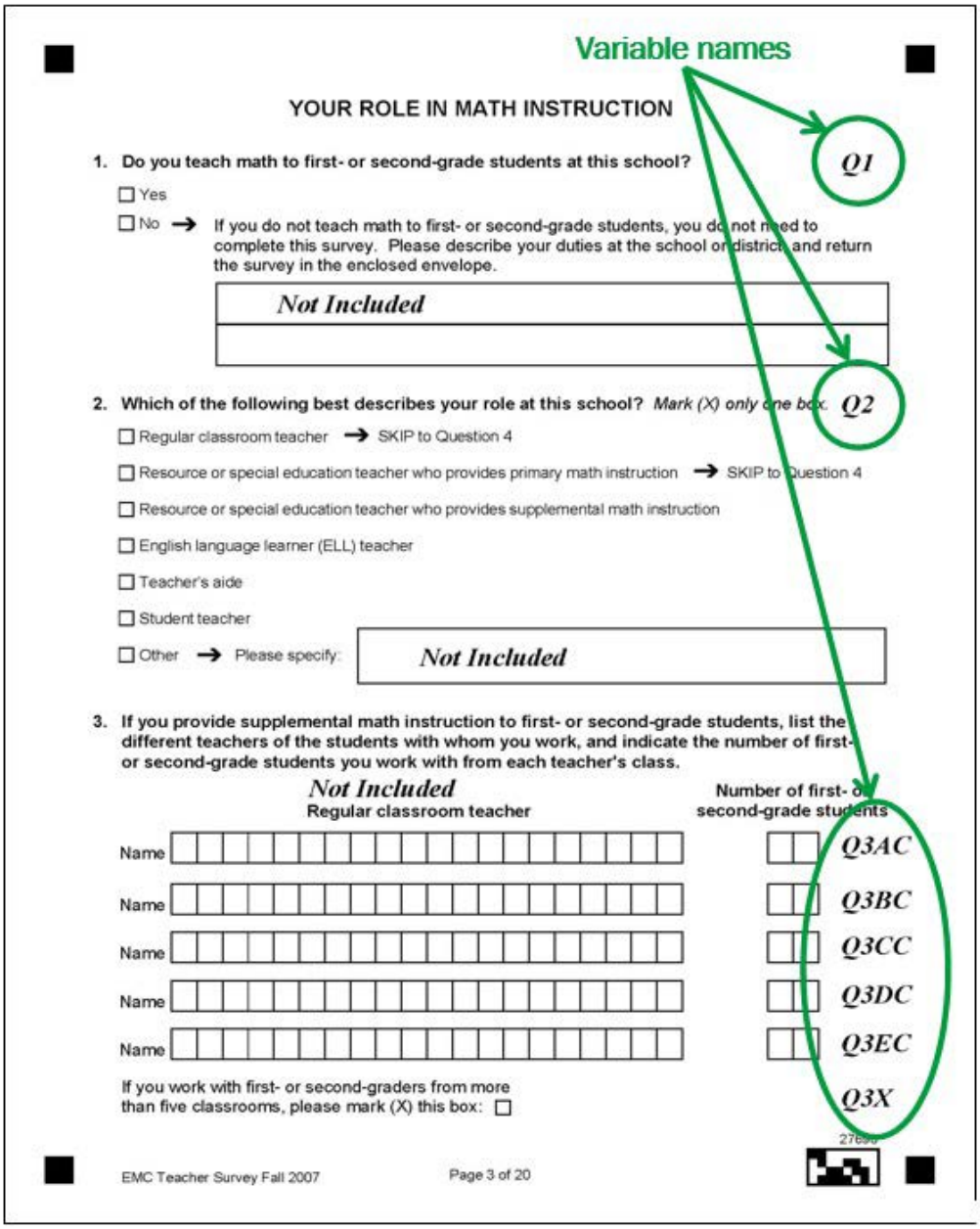


Illustration from Agodini, R. (2015). Guide for constructing a restricted-use data file for NCEE studies. Unpublished manuscript.

Summary statistics document

The document with summary statistics should list variables in their order of appearance in the data instrument(s). Variables containing administrative data should be summarized in the order they appear in the data file. Each variable should be clearly labeled with a variable name and variable label (the survey question or a short description of the variable) so the user can easily identify the corresponding item in the instrument.

Variable name: Q1

Variable label: Do you teach math to first- or second-grade students at this school?

Response	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0=No	19	4.11	19	4.11
1=Yes	443	95.89	462	100.00

Variable name: Q2

Variable label: Which of the following best describes your role at this school?

Response	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1=Regular classroom teacher	428	92.64	428	92.64
4=English language learner (ELL) teacher	6	1.30	434	93.94
7=Other	5	1.08	439	95.02
9999=Refused	23	4.98	462	100.00

Teacher survey question 3. If you provide supplemental math instruction to first- or second-grade students, indicate the number of first- and second-grade students you work with from each teacher's class.

Variable name	Label	N	Mean	Std Dev	Min	Max
Q3AC	No. of students from 1st teacher's class	3	3	.82	2	4
Q3BC	No. of students from 2nd teacher's class	4	2	.71	1	3
Q3CC	No. of students from 3rd teacher's class	3	1.33	.47	1	2
Q3DC	No. of students from 4th teacher's class	3	1.33	.47	1	2
Q3EC	No. of students from 5th teacher's class	3	3	.82	2	4

Variable name: Q3X

Variable label: If you work with first- or second-graders from more than five classrooms, please mark (X) this box.

Response	Frequency	Percent	Cumulative Frequency	Percent
9997=NA	426	.	.	.

Example modeled on Agodini, R. (2015). Guide for constructing a restricted-use data file for NCEE studies. Unpublished manuscript.

Appendix B.4: Sample description of data files

This appendix contains excerpted sections from the Description of Data Files included in a restricted use dataset for an evaluation of alternate routes to teaching published by IES. The study report is available on the IES website (Clark et al., 2013).

Return to resources for documenting and organizing data files and code

Excerpted Sections from The effectiveness of secondary math teachers from Teach for America and Teaching Fellows programs description of data files

Teach For America (TFA) and the TNTP Teaching Fellows programs are important and growing sources of teachers of hard-to-staff subjects in high-poverty schools, but comprehensive evidence of their effectiveness has been limited. To learn about the effectiveness of teachers from these programs, the U.S. Department of Education's Institute of Education Sciences commissioned an evaluation of the impact on student achievement of secondary math teachers from the TFA and Teaching Fellows programs. The evaluation followed an experimental design: students were randomly assigned to a class taught by a teacher from either TFA or a Teaching Fellows program or to a class taught by a comparison teacher who entered the profession through some other route to certification…Mathematica Policy Research, Branch Associates, and Chesapeake Research Associates conducted the study.

The data files that are documented here were used as the basis of the final report, "The Effectiveness of Secondary Math Teachers from Teach For America and Teaching Fellows Programs" (Clark et al. 2013), and were subsequently prepared for release as a set of restricted-use data files. This document describes the structure and content of these restricted-use data files. Section A provides an overview of the study design. Section B discusses the data sources used to compile the data files created for the study. Section C provides information on the datasets included in the restricted use files.

A. Study design

This section describes the study design, including the research questions investigated by the study; the groups of teachers included in the study; the study's experimental design; strategies for recruiting and selecting districts, schools, and teachers into the study; and the procedures by which students were randomly assigned to the participating teachers' classrooms. Further information on the study design can be found in Chapter I, Chapter II, and Appendix A of the study's final report (Clark et al. 2013).

1. Research Questions

This evaluation addresses two primary research questions:

- How effective are TFA teachers at teaching secondary math compared with other teachers teaching the same math courses in the same schools?
- How effective are Teaching Fellows at teaching secondary math compared with other teachers teaching the same math courses in the same schools?

To answer these questions, TFA teachers and Teaching Fellows were each compared with other teachers teaching the same subject in the same school. The study was not designed to compare the effectiveness of TFA teachers with that of Teaching Fellows, because students were generally not randomly assigned between TFA and Teaching Fellows teachers. Teacher effectiveness is measured as the difference in end-of-year math scores between students taught by TFA or Teaching Fellows teachers and comparable students taught by teachers from other routes.

To explore possible reasons for differences in effectiveness, the study also examines differences in characteristics between TFA teachers and comparison teachers and between Teaching Fellows and comparison teachers. These characteristics include demographic characteristics, educational background, teaching and nonteaching work experience, and differences in their knowledge of mathematics. It also documents how TFA and Teaching Fellows programs select, train, place, and support their teachers.

Finally, the study explores the relationship between student math achievement and observable teacher characteristics, such as educational background, teaching experience, and math content knowledge. It also explores the extent to which these characteristics might explain any differences in teacher effectiveness that are found between TFA teachers or Teaching Fellows and teachers from other routes to certification. That section of the study is based on nonexperimental methods and hence is less conclusive than the other components of the study.

2. Types of Teachers in the Study

This study separately compared the effectiveness of TFA and Teaching Fellows teachers with the effectiveness of teachers from other certification routes. Because the effectiveness of the latter group served as the point of comparison for the effectiveness of TFA and Teaching Fellows teachers, we refer to teachers from certification routes other than TFA and the Teaching Fellows programs as comparison teachers. In designing the study and deciding which teachers to include in the sample, our goal was to ensure that the comparison teachers represented a meaningful and appropriate counterfactual—the types of teachers that would have been available had teachers from TFA or the Teaching Fellows programs not been available in a particular school...

3. Experimental Design

To assess the impacts of TFA and Teaching Fellows teachers compared with teachers from other certification routes, the study used an experimental design based on the random assignment of students to teachers. Students in the same school who enrolled in the same math course were randomly assigned to a math class taught by a TFA or Teaching Fellows teacher or to a comparable class taught by a comparison teacher. Random assignment ensured that there were no systematic differences at the start of the school year between students assigned to TFA or Teaching Fellows teachers and those assigned to teachers from other certification routes.

We refer to a group of classes between which students could be randomly assigned as a classroom match. The study included two main sets of classroom matches—those that included TFA and comparison teachers, and those that included Teaching Fellows and comparison teachers. Within both types of matches (TFA or Teaching Fellows), students who were randomly assigned to a TFA or Teaching Fellows teacher constituted the treatment group, and those randomly assigned to a comparison teacher constituted the control group. Classes of TFA teachers or Teaching Fellows and those of comparison teachers in the same school could form an eligible classroom match only if the classes met the following criteria:

- The classes covered a math course that was deemed eligible for the study.
- The classes were similar in subject taught and class conditions.
- The classes were in the same period.[1]

Students in the classes did not receive supplemental math instruction in a way that would systematically inflate or dampen the effects of their regular math teachers.

While most classroom matches consisted of only two classes, some contained more than two. Ninety-three percent of the matches in the TFA study sample and 83 percent in the Teaching Fellows study sample consisted of exactly one class taught by a TFA or Teaching Fellows teacher and one class taught by a comparison teacher. The remaining matches consisted of more than two classes. This could occur if, for example, students were randomly assigned between three comparable math classes—such as three sections of Algebra I taught by two TFA teachers and one comparison teacher—held at the same time.

4. Recruitment of Districts, Schools, and Teachers

We conducted the study over a two-year period and recruited two cohorts of sample members—one that participated in the 2009-2010 school year (cohort 1) and a second that participated in the 2010-2011 school year (cohort 2). A separate sample of districts, schools, and teachers was recruited into the study for each of the two study years, but sample

[1] A few exceptions to this rule are described below.

members could be included in both years. The final sample consisted of 11 states, 15 districts, 82 schools, 228 classroom matches, 287 teachers, and 517 classes.

The sample consisted of two distinct sets of study participants. One set of participants, referred to as the TFA study sample, consisted of TFA and comparison teachers who taught matched classes and included the students who were assigned to those classes and the schools and districts in which those classes were located. The other set of participants, referred to as the Teaching Fellows study sample, consisted of Teaching Fellows and comparison teachers who taught matched classes, along with the corresponding students, schools, and districts. A comparison teacher could belong to both study samples if he or she was compared with both TFA and Teaching Fellows teachers. Table B.4.1 lists the sample sizes for each study sample as well as combined sample sizes.

Table B.4.1: Number of States, Districts, Schools, Classroom Matches, Teachers, and Classes in the Study

	Number of Sample Members in		
	TFA Study Sample	Teaching Fellows Study Sample	Both Samples Combined[a]
States	8	8	11
Districts	11	9	15
Schools	45	44	82
Classroom Matches	111	118	228
Teachers	136	153	287
TFA or Teaching Fellows teachers	66	69	135
Comparison teachers	70	84	152
Classes	248	270	517
Taught by TFA or Teaching Fellows teachers	123	135	258
Taught by comparison teachers	125	135	259

[a] Counts of sample members in the TFA and Teaching Fellows study samples can sum to more than the total count in the combined sample because some comparison sample members belonged to both the TFA and Teaching Fellows study comparison samples. TFA = Teach For America.

Study participants were recruited in similar ways for both the TFA and Teaching Fellows study samples. We focused recruitment efforts on districts with large concentrations of secondary math teachers from TFA or a Teaching Fellows program. Using fall 2008 placement data from TFA and the Teaching Fellows programs, we identified those districts and contacted 42 of them prior to the first study year. Of those, 15 districts agreed to participate in the study by allowing the study team to conduct random assignment and data collection activities.

Within participating districts, we contacted schools in the spring prior to each study year to identify those in which the study could be implemented in the upcoming year. We placed

priority on contacting schools in which TFA and Teaching Fellows programs had previously placed secondary math teachers because those schools had the greatest likelihood of having teachers eligible for the study in the upcoming study year. In each contacted school, we ascertained course schedules and teaching assignments to determine whether the school would have any eligible classroom matches in the following school year. Of the 792 schools that were initially contacted, the final sample of 82 schools consisted of those that contained eligible classroom matches, agreed to allow random assignment of students, and provided verification that students had been placed into classes in accordance with the results of the random assignment.

The final set of classroom matches in the study spanned several types of middle school and high school math courses. Middle school math courses constituted 75 percent of classroom matches with TFA teachers and 31 percent of classroom matches with Teaching Fellows. Of the 287 teachers in the study, about half (135) were either TFA or Teaching Fellows teachers. The number of TFA teachers (66) was similar to the number of Teaching Fellows (69). The total number of teachers in the study (287) was not twice the number of classroom matches (228) because some matches had more than two teachers. Moreover, some teachers in the study taught classes in more than one match—in different periods during the school day and/or in both study years...

5. Selection and Assignment of Students

Before the start of each new school year, schools sent us lists of students whom they wanted placed in one of the classes in the identified classroom matches. We randomly assigned the students in the classes, specifying the teacher for each class. The schools then assigned the students to classes in accordance with the random assignment results. Students who wanted to enter one of the classes after this initial assignment but before the end of the first month of the school year were also randomly assigned to one of the classes. Schools could explicitly request a specific assignment for a given student, in which case the student was excluded from the study (which was rare).

Based on this random assignment, a student was included in the study's research sample—that is, the beginning-of-year sample representing the students to whom study findings pertain—if he or she met two criteria. First, the student was randomly assigned to a study class by the end of the first month of the school year. Second, the student did not leave the school in which the study was being conducted prior to the first day of the school year. All research sample students with valid end-of-year math scores were included in the impact estimates; these students are referred to as the analysis sample. We requested parental consent to obtain data on students; if a student's parents did not give consent for data collection, we did not obtain an end-of-year math score for them and they were excluded from the analysis sample. Table B.4.2 lists the number of students in the research sample and the analysis sample.

Table B.4.2: Number of Students in the Study

Subject	Teach For America Study Sample			Teaching Fellows Study Sample		
	Assigned to TFA Teachers	Assigned to Comparison Teachers	Total	Assigned to Teaching Fellows	Assigned to Comparison Teachers	Total
Number of Students in the Research Sample	2,884	2,906	5,790	3,466	3,443	6,909
Number of Students in the Analysis Sample	2,292	2,281	4,573	2,127	1,989	4,116

Note: 24 students from the research sample–of whom 20 were in the analysis sample–were assigned to a comparison teacher who was in both the Teach For America study sample and the Teaching Fellows study sample. TFA = Teach For America.

There was some student movement into and out of the study classes after the random assignment period. Some research sample members transferred out of their originally assigned classes and some late-enrolling students were placed by schools into study classes after the first month of the school year. Despite this mobility, study classes remained primarily composed of research sample members throughout the year.

B. Data sources

We collected data on students, teachers, schools, and the TFA and Teaching Fellows programs in the study, which are summarized in Table B.4.3 below.

1. Data on Students

We attempted to collect data on math achievement and demographic characteristics for all students in the research sample for whom we received parental consent to collect these data.

Math achievement outcomes. We determined end-of-year math achievement based on different sources of data for middle school and high school students in the study. For students in grades 6 to 8, we obtained scores on state assessments from district administrative records.

For students in grades 9 to 12, we administered end-of-course math assessments developed by the Northwest Evaluation Association (NWEA)...

Table B.4.3: Data Sources for the Evaluation

Domain	Data Source	Schedule of Data Collection	
		Cohort 1 (2009–2010)	Cohort 2 (2010–2011)
Student Math Achievement Outcomes			
Middle school	District administrative records	Summer/Fall 2010	Summer/Fall 2011
High school	NWEA assessments (study-administered)	Spring 2010[a]	Spring 2011[b]
Baseline Student Achievement and Characteristics	District administrative records	Summer/Fall 2010	Summer/Fall 2011
Student Mobility	Class rosters	Summer 2009, Fall 2009, Winter 2010, Spring 2010	Summer 2010, Fall 2010, Winter 2011, Spring 2011
Teachers' Route to Certification	Teacher background form	Summer/Fall 2009	Summer/Fall 2010
Teachers' Professional Background and Experiences	Teacher survey	Spring 2010	Spring 2011
Teachers' Math Content Knowledge	Praxis (study-administered or existing score from Educational Testing Service)	Fall 2009	Fall 2010
School Characteristics	Common Core of Data	2009–2010	2009-2010
TFA and Teaching Fellows Program Characteristics	Program administrator interviews	Spring 2010	n.a.

[a] Assessments for single-semester fall 2009 classes were conducted in winter 2009–2010.

[b] Assessments for single-semester fall 2010 classes were conducted in winter 2010–2011.

TFA = Teach For America; NWEA = Northwest Evaluation Association.

n.a. = not applicable.

Students in the research sample with valid outcome scores constituted the final sample used in the analysis. A student may not have had a valid outcome score for one of four reasons: (1) the student's parents did not give consent for data collection; (2) we were not able to administer an NWEA assessment to the student (if in high school); (3) the student's score on the NWEA assessment was invalid based on criteria defined by NWEA (see Appendix A of Clark et al. 2013); or (4) the district did not have a state assessment score for the student (if in middle school). We obtained valid outcome scores for 68 percent of students in the research sample—in particular, for 70 percent of treatment group students and 67 percent of control group students. On average, we had valid outcome scores for 79 percent of students in TFA classroom matches and 60 percent of students in Teaching Fellows classroom matches. Additional information on percentages of the sample with valid outcome scores can be found in Appendix A of Clark et al. 2013...

2. Data on Teachers

Route to certification. Before each year of the study, we verified the certification route of all teachers whose classes could potentially be included in classroom matches by asking principals to complete a background form on each study teacher…

3. Data on Schools

…

4. Data on TFA and Teaching Fellows Programs

…

C. Data files

Ten data files are included in the restricted-use file package. The student, teacher, classroom, and school analysis files each contain data from multiple sources that were cleaned, processed, and used to produce the study results. The student imputation files contain data calculated from the information in the student analysis file, which were then used to fill in missing information for the student analysis. The teacher survey and program interview files contain data collected from these two sources. (The full data collected from other sources, such as the class rosters and teacher background forms, are not included in order to protect the confidentiality of study participants. All data from these sources used in the report are included in the analysis files.)

As mentioned previously, the Teaching Fellows programs in this study are run by TNTP, a national nonprofit organization, in partnership with school districts. Each data file uses the term "TNTP" or "tntp" in variable names, labels, and other documentation to refer to the Teaching Fellows sample. The corresponding term for the Teach For America sample is "TFA" or "tfa."

Table B.4.4 lists the name, record level, sample size, and number of variables of each dataset included in the restricted-use files.

The next seven sections contain detailed information about each data file. The four multiply imputed files are reviewed together in section 2. Information that is common to all files is then discussed in section 7.

1. Student Analysis File

Sample definition. The student analysis file contains 21,144 observations—one for each student who was ever assigned to a classroom in which the study team conducted random assignment. Of these students, 12,675 students constitute the research sample (defined by the variable rabase_large)—the beginning-of-year sample representing the students to whom study findings pertain. Students were included in the research sample only if they met several conditions. First, they needed to have been assigned to a study class in which the school implemented the results of the study's random assignment process. Classes in which the

school failed to implement the random assignment results are flagged by the variable classdropped…Third, the students could not have left their original school before the start of the school year. Students who left before the start of the school year are flagged with a value of 1 for either flag_df3 or flag_dt3…

Table B.4.4: Datasets Included in the Restricted-Use Files

Dataset	File Name	Record Level	N	# of Variables
Analysis Files				
Student Analysis File	student_analysis.dta	Student	21,144	287
Multiply Imputed Files				
	student_imputed_tfa.dta	Student	5,462	211
	student_imputed_tntp.dta	Student	5,313	211
	stu_tchr_imputed_pooled.dta	Student	8,689	722
	stu_tchr_imputed_pooled_xsm.dta	Student	8,689	722
Teacher Analysis File	teacher_analysis.dta	Teacher	323	207
Classroom Analysis File	classroom_analysis.dta	Classroom-Teacher	523	24
School Analysis File	school_analysis.dta	School	63,148	28
Raw Data Files				
Teacher Survey	teacher_survey.dta	Teacher	301	227
Program Interviews	program_interviews.xlsx	Program	20	106

Within the research sample, 8,669 students constitute the analysis sample (defined by the variable inexanalys)—the sample used for the study's impact analyses. In order to be included in the analysis sample, students needed to have a valid end-of-year math score from the study school year. Within the research sample, five groups of students were excluded from the analysis sample: (1) students without parental consent to have data collected for the study (defined by the variable consent_final)…Random assignment units are defined in detail below…

Student, school, and district identification variables. The variable student_id uniquely identifies students… The variables sch_id and dist_id uniquely identify the participating school and district in which a student was enrolled when the student entered the study.

Random assignment variables. Several variables contain information about the methods and results of random assignment for each student. The variable treat indicates whether each student was initially assigned to the treatment group (the classroom of a TFA or Teaching Fellows teacher) or the control group (the classroom of a comparison teacher) when the student first enrolled in a study classroom…

Random assignment was conducted separately within each random assignment unit (or, synonymously, classroom match)—a group of comparable classrooms between which students

were randomly assigned. Within each random assignment unit, random assignment was conducted separately within strata defined by date of assignment and up to three types of student-level characteristics selected by the participating schools. The variable stratumid uniquely identifies these strata…

The file also includes variables related to analysis weights that reflect assignment probabilities. For each student who was randomly assigned to a study classroom, the variable raprob indicates the probability of being assigned to the treatment status (treatment or control) to which the student was actually assigned…

Mobility information…[T]he file contains several variables that identify the classrooms in which a student was enrolled at various points in time during the school year. The variable origclass_id identifies the ID of the classroom to which each student was originally assigned when the student entered the study…

Outcome math scores, expressed as z-scores, are contained in the variables mpostzpub and mpostzsam. The z-scores in mpostzpub were standardized using means and standard deviations of scores in a statewide or national reference population. For a student's score on a state assessment, the reference population was the full population of students who took the same assessment in the same state, year, and grade…

Variables with imputed values. The variables described thus far do not contain imputed values. However, the main analyses in the study report used mean imputation to fill in missing values of covariates in the estimation models—in particular, variables measuring student background characteristics and baseline achievement (Clark et al. 2013, Appendix B). For each original variable used as a covariate, there is a corresponding variable that contains mean-imputed values (for students who had missing values of the original variable) and original values (for students who had nonmissing values of the original variable). The name of the latter variable is the name of the original variable suffixed with _ii…

2. Multiply Imputed Files

Various analyses in the study report used multiple imputation. First, in sensitivity analyses, impacts were estimated based on data in which missing values for either the covariates or the outcome variable (or both) were replaced with multiply imputed values. The multiply imputed data used in these sensitivity analyses are in the TFA impact analysis imputation file (student_imputed_tfa.dta) and Teaching Fellows impact analysis imputation file (student_imputed_tntp.dta)…

3. Teacher Analysis File

…

4. Classroom Analysis File

…

5. School Analysis File

…

6. Teacher Survey

Sample definition. This file contains an observation for each teacher who completed the teacher survey, for each cohort in which he or she completed it. We attempted to administer the teacher survey to all teachers in the study (both original and replacement teachers). For teachers who were in the study in both cohorts, we asked them to complete the full survey in cohort 1 and a follow-up survey in cohort 2. If a teacher was in both cohorts but did not complete the survey in the first year, we asked him or her to complete the full survey in the second year.

Observations corresponding to teachers who responded only to the full cohort 1 or 2 survey have the variable surveynum equal one or two, respectively…

7. Common Data File Information

Merging datasets. The student analysis file and the four multiply imputed files can all be merged together using the variable student_id. The TFA and Teaching Fellows impact analysis imputation files were designed to be merged with the student analysis in order to conduct additional analyses, as described above in section 2…

The school analysis file can be merged onto other files that have school IDs by using the variable sch_id. Only information for study schools can be merged with other files; observations for non-study schools cannot be merged with any other file…

Missing values. Table B.4.5 below lists the missing codes used in the restricted-use files…

Table B.4.5: Types of Missing Data Values

Value (Stata)	Meaning	Applicable Datasets
.l	Logical skip	Teacher Survey; Program Interviews
.n	Not applicable	Teacher Survey; Program Interviews; School Analysis
.m	Missing with no other code applied	All

Restricted-use file package. The CD includes two folders: (1) Datasets and (2) Documentation. The data folder contains the Stata format files for the datasets discussed above in sections 1 through 7…

The Documentation folder contains PDF codebooks and files with summary statistics for each source. Table B.4.6 shows the complete list of released files.

Table B.4.6: Organization of Restricted-Use Files

Description of Data Files.pdf
Datasets (folder)
student_analysis.dta
student_imputed_tfa.dta
student_imputed_tntp.dta
student_tchr_imputed_pooled.dta
student_tchr_imputed_pooled_xsm.dta
teacher_analysis.dta
classroom_analysis.dta
school_analysis.dta
teacher_survey.dta
program_interviews.xlsx
Documentation (folder)
all_codebooks.pdf
student_analysis_summstats.pdf
student_imputed_tfa_summstats.pdf
student_imputed_tntp_summstats.pdf