Data Sharing and Data Shared

Preconference Workshop for SSSR 2021

Jessica Logan, PhD

The Ohio State University

Logan.251@osu.edu | @Jarlogan

*Some of this talk is based on:

Logan, J. A., Hart, S. A., & Schatschneider, C. (2021). Data sharing in education science. *AERA Open*, *7*.

Data Sharing

- Most federal grants now require you to share your data
  - Required by IES, NIH, and NSF.

- Data sharing has been a requirement for IES Goal 4 / 3 since FY2013 / FY2014
  - About 350 research projects funded by IES are subject to this data sharing requirement, but few of them have shared the data (Albro, 2020)

- Education researchers believe data sharing to be the least common of the open science practices (Makel et al., 2019)

Data Sharing: What is it?

- Taking any type of research data and making it available for others to examine or use.
  - Reporting means, SDs, and correlation matrices is a form of data sharing.
  - This means you've almost certainly already shared *some* data!

- More formally, "data sharing" typically refers to the sharing of participant- and variable-level data, not simply summary statistics

Why to share?

- It supports the Scientific Process
  - Data collected by researchers can be biased towards proposed hypotheses (Wicherts et al., 2016).
  - Different analytic choices can show different results, even within the same dataset and using the same variables (Silberzahn et al. 2018)

- Greater transparency in science
  - Other researchers can check your published work.
  - Others can replicate or extend analyses that you have already published.

Why to share?

- Opens your resources to others.
  - Helpful to researchers who are at early stages of their careers, or who are working at institutions or organizations that do not have the infrastructure to support large research projects.

- Enables studying low occurrence samples:
  - Integrative data analysis can combine multiple datasets (Curran & Hussong, 2009).
  - For people studying due to low numbers of participants or low occurrences of a given behavior of interest.

Why? Benefits the Depositor

- Datasets that you share get a DOI and are citable
  - Can be included in a measure of your research impact

- Papers with open access to data have an increased citation rate

  (Drachen, Ellegaard, Larsen, & Dorch, 2016; Piwowar et al., 2007; Piwowar & Vision, 2013).

- Papers with data in a repository (without restriction) had up to a 25% higher citation rate (Colavizza et al., 2020).

Principles of Data Sharing

- There are four internationally accepted features of good data sharing, called the FAIR data principles*:

  - Findable: Datasets should be stored with rich metadata that describes what's inside the file without opening it.

  - Accessible: Anyone with a computer and internet should be able to access at least descriptive information about the study and dataset (the metadata).

  - Interoperable: Data and metadata need to be stored so other computers can read it.

  - Reusable: Anyone who does find and access your data also needs to have all the critical information that will help them understand your data, and people know how to cite it.

*(Wilkinson, 2016)

How To Share

- Steps to take to share data that meets the four FAIR standards

- For data to be reusable, you need to provide lots of Metadata along with the actual data files.

To share your data, you need to include:

1. Study Description
    a) Study summary
    b) Protocols for design, methodology, procedures

2. The Prepared Data File (or Files)
    a) Submit cleaned and prepared file
    b) Must be de-identified

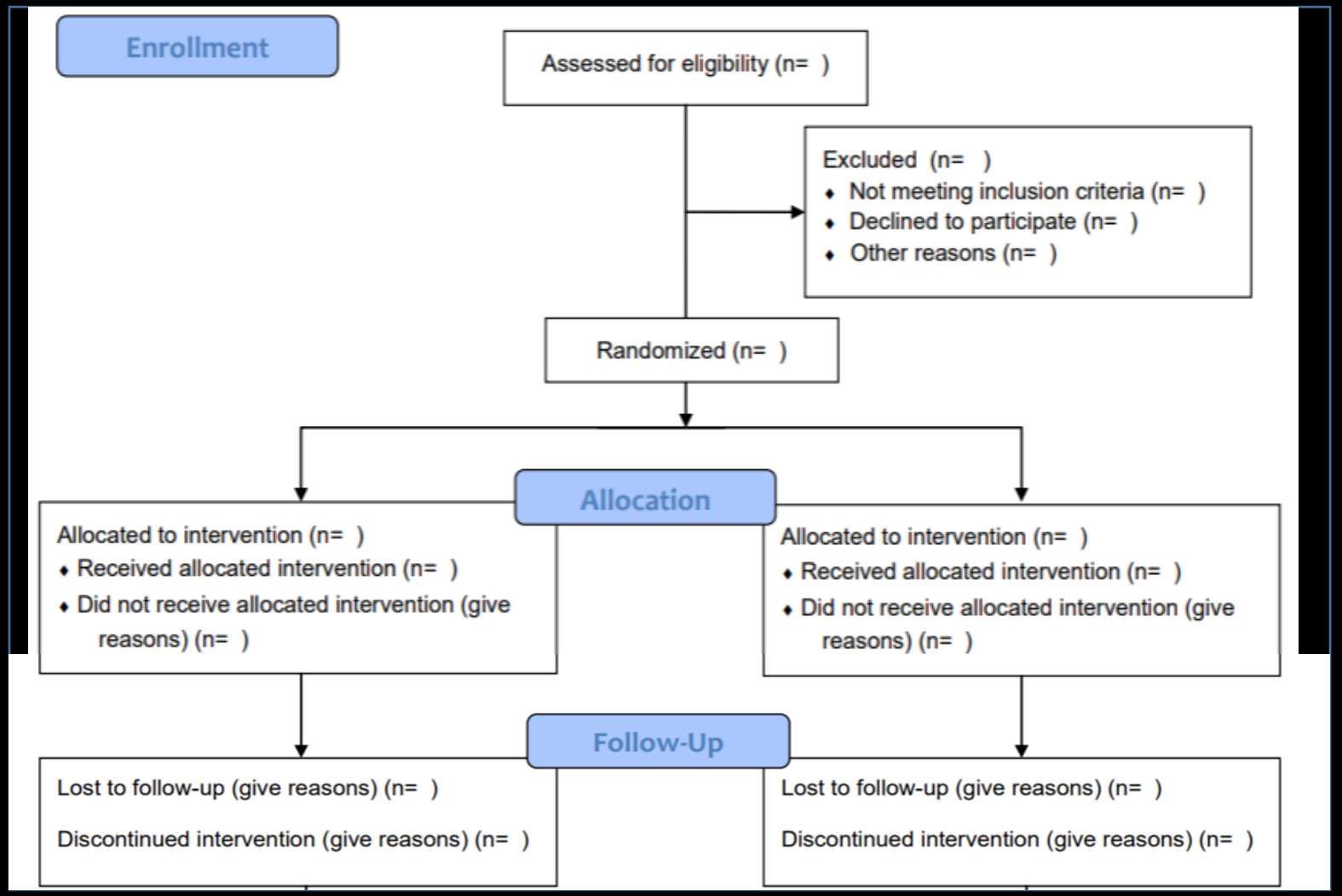
3. Data File Documentation
    a) Codebooks / Variable Names / data dictionary
    b) Documentation of missing codes

1a. Study Summary

- Provide the overarching view of the study.

- Help the potential user understand your study motivations:
  - Background and research questions
  - Description of the research design
  - Random assignment strategy (if applicable)
  - Recruitment and data collection procedures
  - Sample retention (try a CONSORT diagram; consort-statement.org)

Consort Diagram



1b. Protocols

- Protocols are written documentation of project decisions.
  - They are living documents and should be frequently updated
  - When your project is done, compile these into one explanatory document and submit with your file.
- Protocols document what decision was made and why.
  - The decision is a rule to be implemented across the project.
  - Make the decision. Write it down. Don't change it.
- Future you won't remember what current you decid
  - Or why you decided to do it.

1b. Examples of protocol topics

- Make a decision. Write it down. Don’t change it.
  - Eligibility Criteria
  - Inclusionary criteria
  - Assessment administration
  - ID schemes
  - Establishing observer reliability
  - Data cleaning decisions
  - Variable Naming protocols
  - Missing data codes

Protocol example

RE: *Direct Child Assessment Protocol*

Authored Date: *3/18/2019*

Decision:

The direct assessments will be given in the following order: A) Peabody Picture Vocabulary Test then B) Head Toes Knees Shoulders (HTKS)

Year 1 Time 1 testing window will be 9/22/2019 – 11/1/2019. Students not able to be assessed during this window will be documented as missing (See missing data protocol).

Rationale:

Measures must be given in a consistent order to ensure that we can examine questions of individual differences. The HTKS is more engaging for students so we put it second.

To share your data, you need to include:

1. Study Description
    a) Study summary
    b) Protocols for design, methodology, procedures

2. The Prepared Data File (or Files)
    a) Submit cleaned and prepared file
    b) Must be de-identified

3. Data File Documentation
    a) Codebooks / Variable Names / data dictionary
    b) Documentation of missing codes

2. The Prepared Data:

- To clean and prepare data for sharing, must decide what to include in your data files:

- Should you include all of the variables?
  - Probably not

- Think about:
  - Can you include item-level information?
  - Making sure the file is de-identified

Decide WHAT to share:
Can you share items?

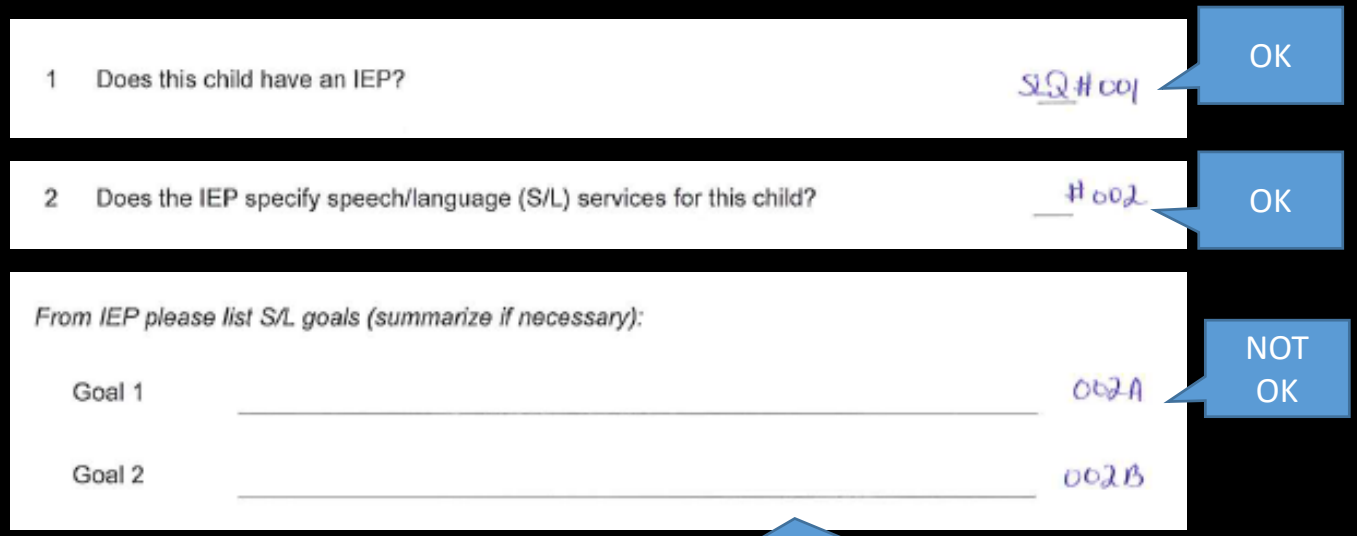
- Published scales
  - Copyright may prevent you from including a complete scale in your documentation
  - You can include item-level information for freely-available scales
  - You can include it if item-specific information is excluded:
    - OK: Item 1 of the Woodcock Johnson III Picture Vocabulary
    - Not OK: “Bottle”

Decide WHAT to share:
De-Identification

- Create new ID variables
  - Must be independent of in-house data collection procedures.
  - Can’t be traceable to identifiable information.

- To make data de-identified, remove:
  - Names of respondents, buildings, or organizations.
  - Birthdates and Test Dates (convert into age in months at testing)
  - Address information (State is OK, but nothing else)
  - Text fields / responses to open ended questions
    - e.g., “Juan struggles with his ST sounds”.

Decide what to share: Case



OK

OK

NOT OK

Options for questions like this:

1. Redact any identifiable information. From each response
2. Recode this data for particular response types
   - (e.g., this goal mentions vocabulary).
3. Completely redact and do not share these items.
4. Share, but require a higher level of security for access

To share your data, you need to include:

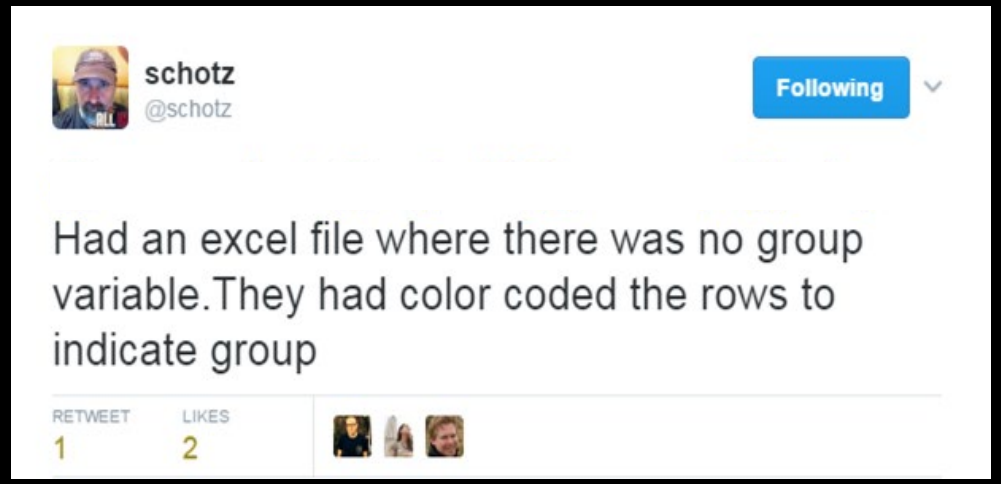
1. Study Description
    a) Study summary
    b) Protocols for design, methodology, procedures

2. The Prepared Data File (or Files)
    a) Submit cleaned and prepared file
    b) Must be de-identified

3. Data File Documentation
    a) Codebooks / Variable Names / data dictionary
    b) Documentation of missing codes

3. Data Documentation

- All data must be stored with letters and numbers
  - Not color or highlighting
  - Remember, machine readable!



3. Data Documentation

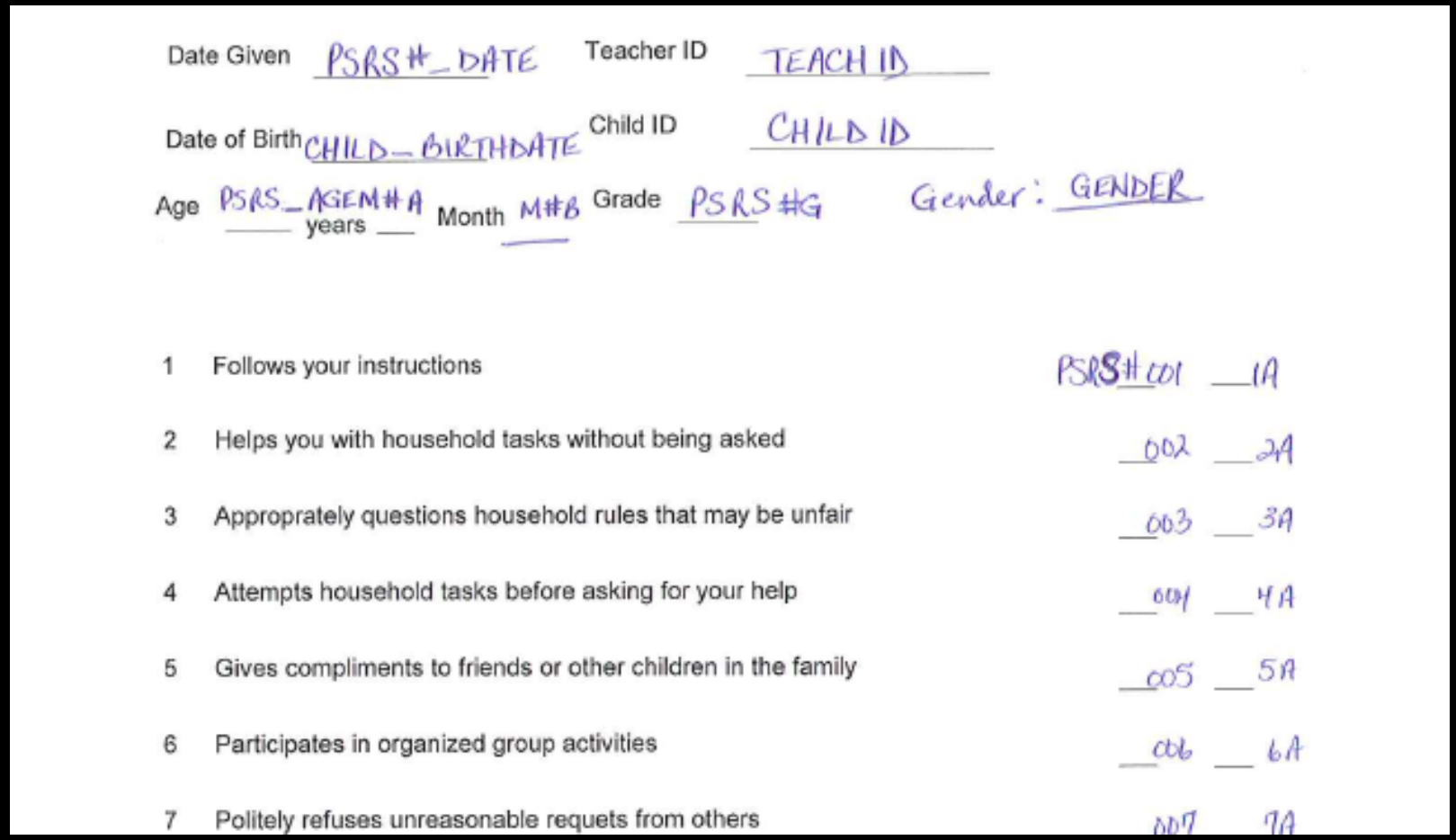
- Variable Names
  - The shortened digital representation of each variable in your dataset
  - Variable names must be unique (no two variables share a name)
  - Different variables have different names

- Develop a rule for how variables will be named.
  - Write it down. Don’t change it.
  - Capture variable naming rules in a protocol.

- Capture variable names in a Variable Name Book or Codebook
  - Record variable information in a Data Dictionary

3. Data Documentation:
Variable Name Book Example

- Image of the questions asked
- With variable names written in the blanks
- This can also be aggregated into a data dictionary:



Date Given PSRS#_DATE Teacher ID TEACH ID

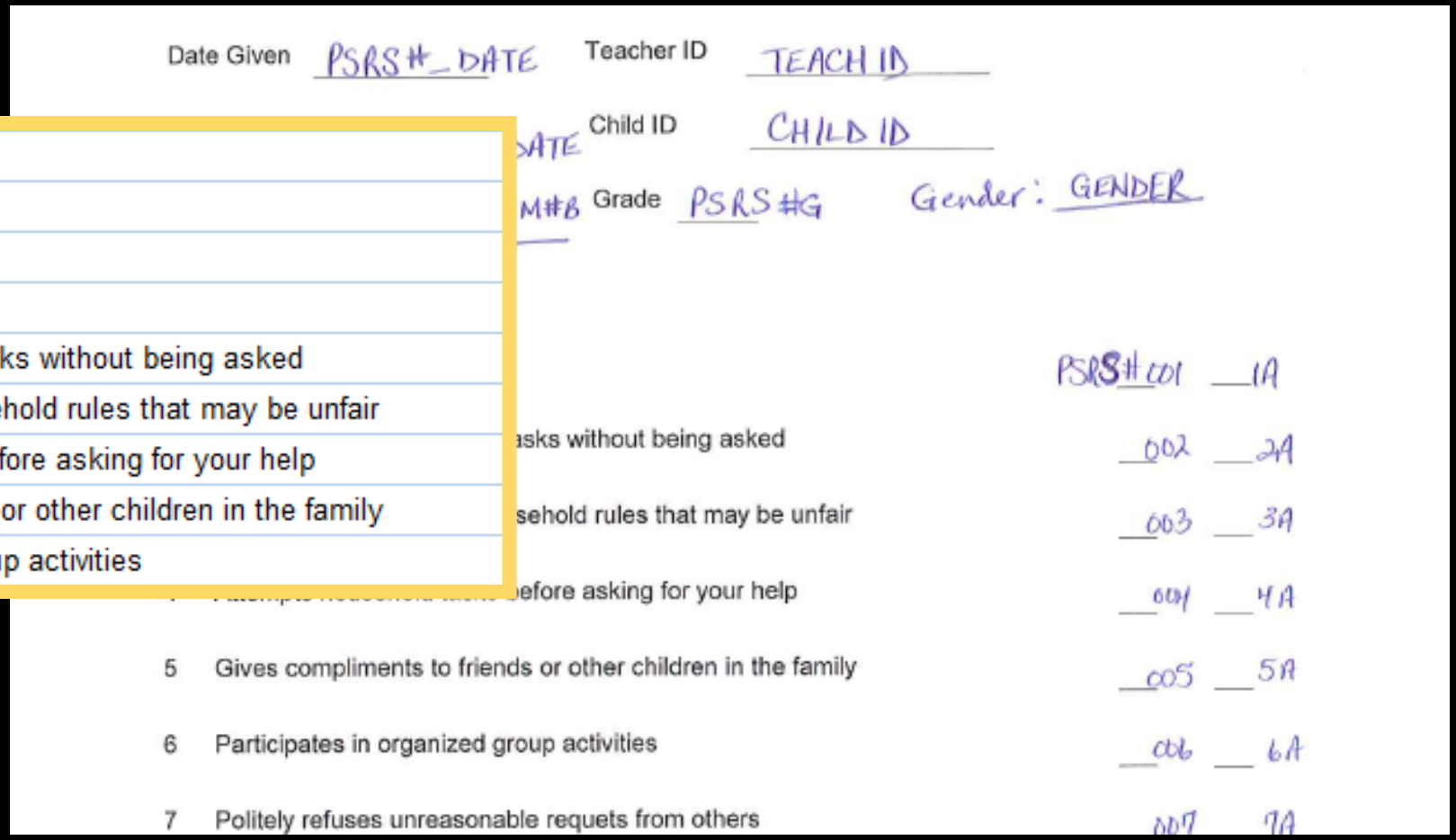
Date of Birth CHILD_BIRTHDATE Child ID CHILD ID

Age PSRS_AGEM#A ____ years ___ Month M#B Grade PSRS#G Gender: GENDER

1	Follows your instructions	PSRS#001	1A
2	Helps you with household tasks without being asked	002	2A
3	Appropriately questions household rules that may be unfair	003	3A
4	Attempts household tasks before asking for your help	004	4A
5	Gives compliments to friends or other children in the family	005	5A
6	Participates in organized group activities	006	6A
7	Politely refuses unreasonable requets from others	007	7A

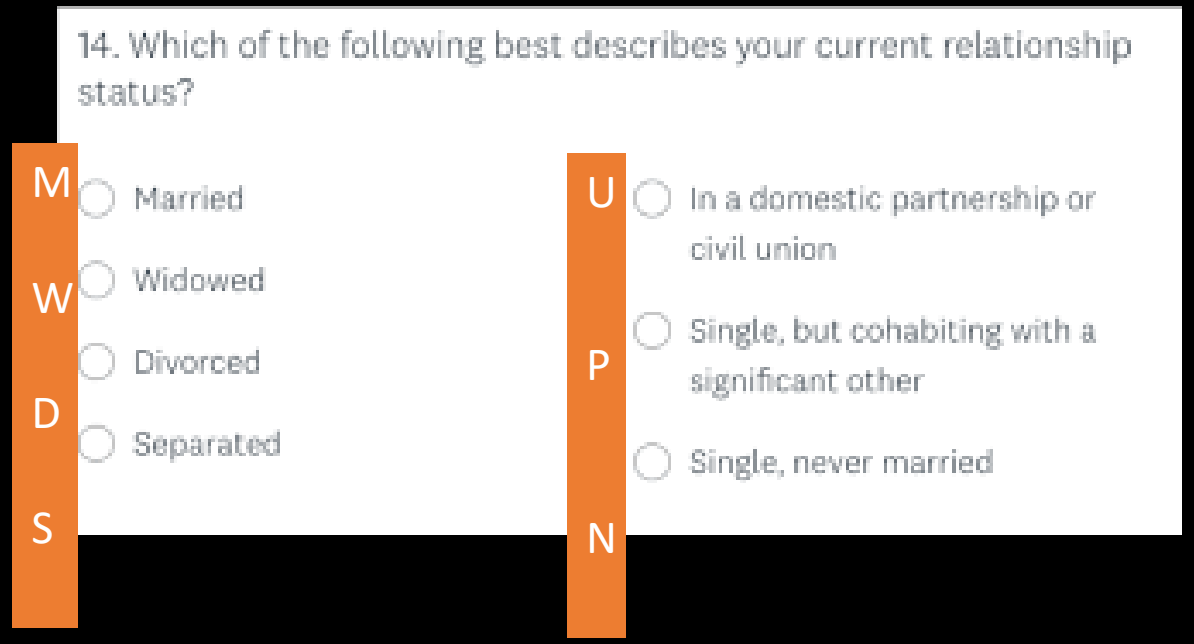
Data dictionary:

PSRS3_DATE	Date Given
PSRS_AGEM3A	Age
PSRS3G	Grade
PSRS3001	Follows your instructions
PSRS3002	Helps you with household tasks without being asked
PSRS3003	Appropriately questions household rules that may be unfair
PSRS3004	Attempts household tasks before asking for your help
PSRS3005	Gives compliments to friends or other children in the family
PSRS3006	Participates in organized group activities



3. Data Documentation

- Variable values are how the responses from participants are stored as numbers.
- How to store this?



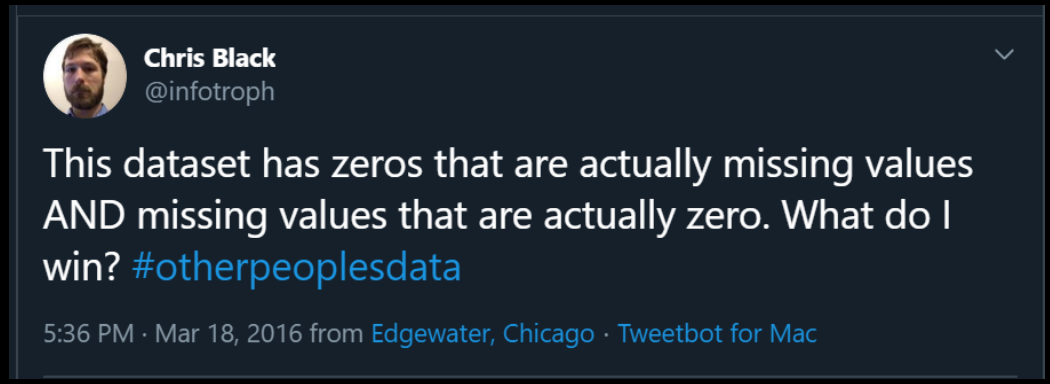
3. Data Documentation

- Variable values are how the responses from participants are stored as numbers.
- How to store this?

- Variable values follow similar rules to variable names.
  - Must be unique
  - Suggest keeping them consistent across your project.

3. Data Documentation

- Document how Missingness is captured in your data.
- Options:
  - Large implausible negative numbers (e.g., -999)
    - Problematic if not clearly documented; must be changed before analysis.
  - As N/A *(in some programs)
    - Problematic in other programs; converts information to character
  - Leaving blank
  - Don’t store them as zero.



Findable and Accessible

“Upon Request”

- You may be familiar with “data are available on request” disclaimer
  - These technically meet most data sharing requirements

- Not very *Accessible*:
  - Historically, many authors have been unresponsive to requests (Wicherts et al., 2006).
  - Data accessibility declines over time

- Not very *Findable*.
  - Know someone who is involved in the project
  - Read about it in a paper

“Upon Request”

- Also has several disadvantages for the data holder:
  - Has to spend time and resources processing any request.
  - Preparing data to share
  - Answering questions from the requester
  - Looking into new rules and regulations (e.g., execute a new data sharing agreement)

- Newer federal guidelines state they prefer a data repository
  - (NIH, 2020; effective 2023)

Data Repository

- A data repository is any place where you can store your data and accompanying metadata that provides access to others.
  - Findable: Many different datasets stored in one place
  - Findable: Data are stored with metadata and are searchable
  - Accessible: By definition, repositories provide access to others.
  - Accessible: Keeps access available for the future (~100 years)
  - Interoperable: Data are stored in forms that can be opened by multiple programs
  - Reusable: Provides citation information

- Here are two different repositories that would be great for the types of work here at SSSR:
  - Note that many exist, see Logan et al., 2021 for a more comprehensive list

Where to share option 1: ICPSR

- Your dataset(s) are posted as a web page
  Findable, Searchable, discoverable online

Where to share option 1: ICPSR

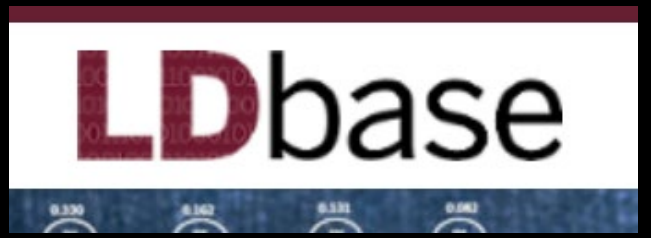
- ICPSR is a repository for datasets from individual projects

- Restrictions and confidentiality:
  - You can register your dataset at five different levels of restricted access
  - Most restricted: can have people go through approval process for access
  - They *offer a* review your project to ensure human subjects confidentiality

- Many datasets that you may be familiar with are stored there
  - ECLS-K, NLSY, Head Start Impact Study

- *It is NOT free!*
  - Time from upload to availability: about 4 months.

Where to share option 2: LDBase



- LDBase is a repository for datasets within projects
  - Created specifically to house student achievement and learning disabilities data

Note: All data is provided in long format, and as such each participant has multiple rows.

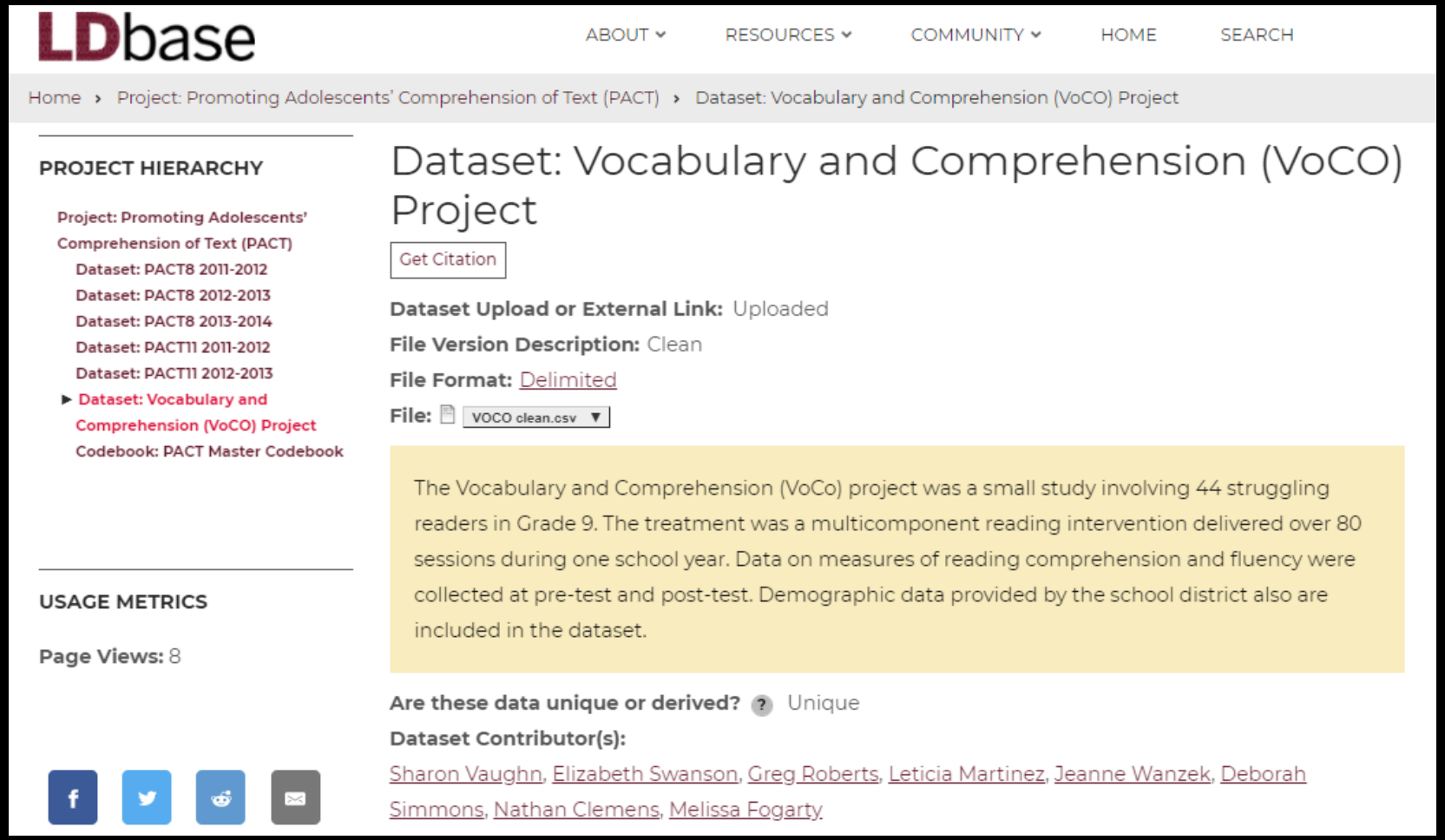
- Restrictions and Confidentiality
  - Multiple restriction and embargo options
  - You review data for confidentiality yourself before submitting.

- Already includes data from 6 NICHD- & IES-funded projects
  - Includes data from ~20,000 children tested longitudinally
  - Now open for you to deposit data

- *It* IS *free!*
  - Time from upload to availability: Instant.

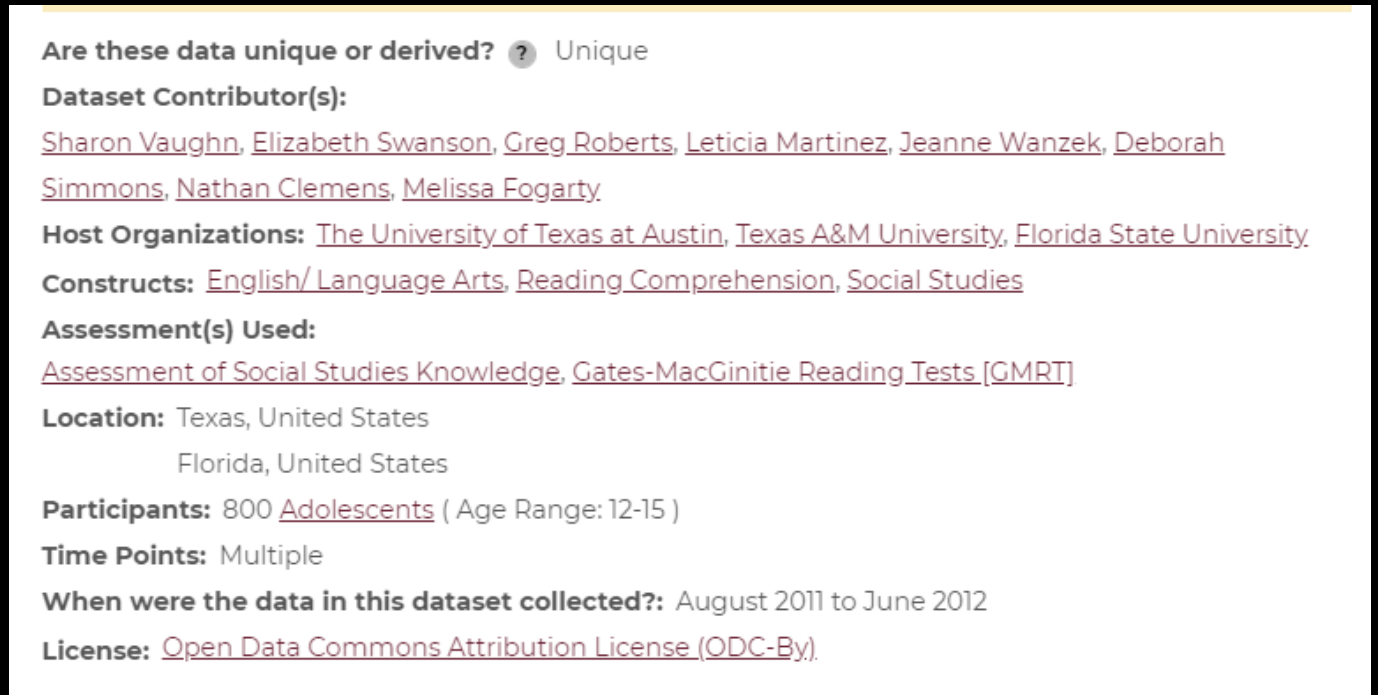
Where to share option 2: LDBase

- Projects get their own web page with datasets nested under each project:



Where to share option 2: LDBase

- Specifically designed for Integrative Data Analysis
  - Depositors have the option to rename variables on deposit to match a master variable name codes
    - Done with a Shiny App online
  - Allows for data to easily be compare across projects.



Data Shared: Using a Data Repository

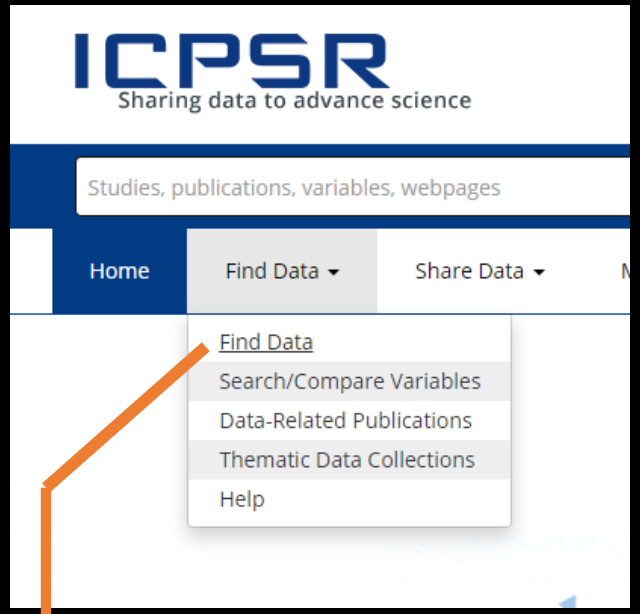
- You can use available data for your own research purposes

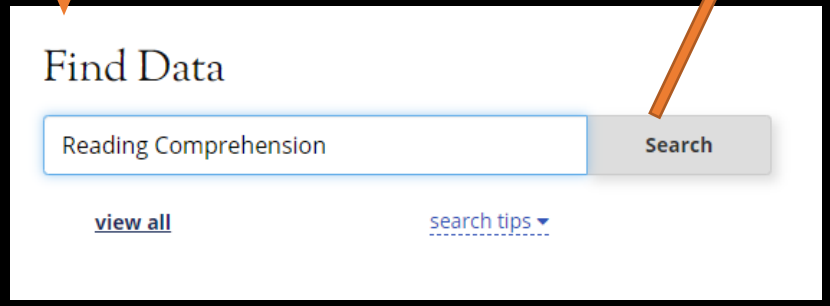
- How do you find data to analyze for a project?

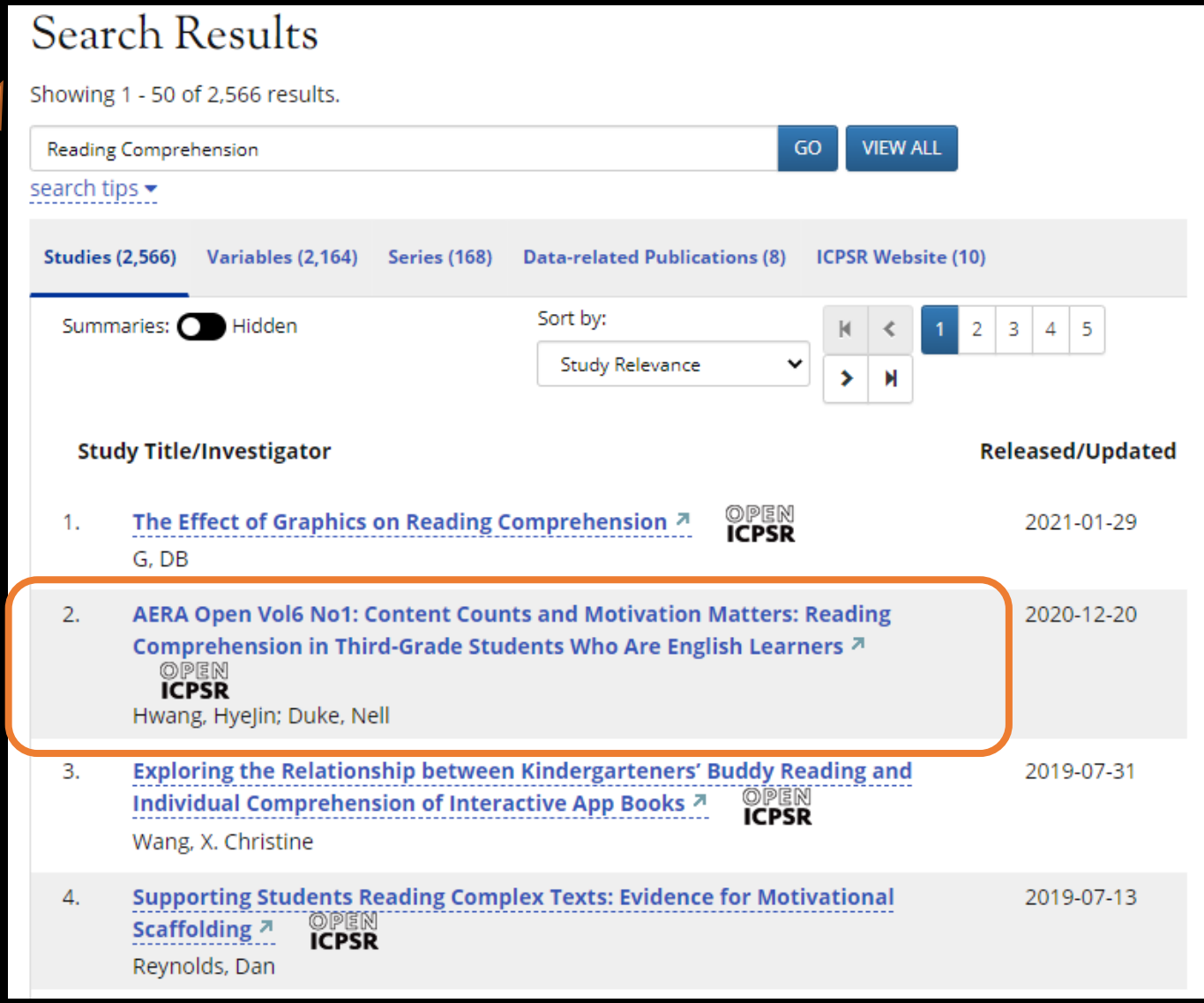
Steps for data reuse

1) Find a dataset (more in a moment)
2) Read all documentation
   - Start with overall project description
   - Learn objective of the study, details of who the sample is
   - Consider any historical events that may influence the data (ahem, 2020)
3) Find at least one previous publication using the data.
   - Read and familiarize yourself with it
   - How is your idea or project different?
4) Download data
   - Try to recreate any provided descriptive or inferential statistics

ICPSR


ICPSR
Sharing data to advance science
Studies, publications, variables, webpages
Home
Find Data
Share Data
Find Data
Search/Compare Variables
Data-Related Publications
Thematic Data Collections
Help


Find Data
Reading Comprehension
Search
view all
search tips


Search Results
Showing 1 - 50 of 2,566 results.
Reading Comprehension
GO
VIEW ALL
search tips
Studies (2,566)
Variables (2,164)
Series (168)
Data-related Publications (8)
ICPSR Website (10)
Summaries: Hidden
Sort by:
Study Relevance
1 2 3 4 5
Study Title/Investigator
Released/Updated
1. The Effect of Graphics on Reading Comprehension
G, DB
2021-01-29
2. AERA Open Vol6 No1: Content Counts and Motivation Matters: Reading Comprehension in Third-Grade Students Who Are English Learners
Hwang, HyeJin; Duke, Nell
2020-12-20
3. Exploring the Relationship between Kindergarteners' Buddy Reading and Individual Comprehension of Interactive App Books
Wang, X. Christine
2019-07-31
4. Supporting Students Reading Complex Texts: Evidence for Motivational Scaffolding
Reynolds, Dan
2019-07-13

AERA Open Vol6 No1: Content Counts and Motivation Matters: Reading Comprehension in Third-Grade Students Who Are English Learners

Principal Investigator(s): HyeJin Hwang, Florida State University; Nell Duke, University of Michigan

Version: V2

Name	File Type	Size	Last Modified
AERAOpen_Vol6_No1.do	text/x-stata-syntax	4.4 KB	12/20/2020 04:36:PM
AERAOpen_Vol6_No1_Chi-square.inp	text/plain	1.3 KB	12/20/2020 03:59:PM
AERAOpen_Vol6_No1_Correlation.inp	text/plain	1 KB	12/20/2020 04:36:PM
AERAOpen_Vol6_No1_Imputed Data.zip	application/zi	10.5 MB	12/20/2020 04:37:PM
AERAOpen_Vol6_No1_Regression.inp	text/plain	1.3 KB	12/20/2020 03:57:PM
README.docx	application/vn officedocum	3.4 KB	12/20/2020 04:44:PM

Data files

Read any explainer documents first.

Metrics for views, downloads, and citations

Usage Metrics

Overall Project Metrics

202	6	1
Views	Downloads	Publications

Download Detailed Metrics

Published Versions

V2 [2020-12-20]

V1 [2020-12-20]

Can upload new versions

Project Citation:

Hwang, HyeJin, and Duke, Nell. AERA Open Vol6 No1: Content Counts and Motivation Matters: Reading Comprehension in Third-Grade Students Who Are English Learners. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-12-20. https://doi.org/10.3886/E129401V2

How to cite the data if you use it.

Project Description

Summary:

This study examined the role of science domain knowledge, reading motivation, and decoding skills in reading comprehension achievement in third-grade students who are English learners (ELs) and students who are monolingual, using a nationally representative data set. Multigroup probit regression analyses showed that third-grade science domain knowledge and motivation for reading, decoding skills, and early attainment of decoding skills were significantly associated with third-grade reading comprehension in both language groups. Also, using Wald chi-square tests, the study showed that the association between third-grade science domain knowledge and reading comprehension was stronger in students who were ELs than in students who were monolingual. These findings suggest that cultivating science domain knowledge is very important to supporting reading comprehension development in third grade, particularly for students who are ELs.

The corresponding publication. There can be more than one!

Related Publications

The following publications are supplemented by the data in this project.

- Hwang, HyeJin, and Nell K. Duke. "Content Counts and Motivation Matters: Reading Comprehension in Third-Grade Students Who Are English Learners." *AERA Open* 6, no. 1 (January 2020): 1–17. https://doi.org/10.1177/2332858419899075.

Export Metadata

Dublin Core

DDI 2.5

Copyright information

Back to search results

Search Results

Showing 1 - 50 of 2,566 results.

Reading Comprehension | GO | VIEW ALL

search tips

Studies (2,566) | Variables (2,164) | Series (168) | Data-related Publications (8) | ICPSR Website (10)

Summaries: Hidden

Sort by: Study Relevance

1 2 3 4 5

	Study Title/Investigator	Released/Updated
1.	The Effect of Graphics on Reading Comprehension (OPEN ICPSR) G, DB	2021-01-29
2.	AERA Open Vol6 No1: Content Counts and Motivation Matters: Reading Comprehension in Third-Grade Students Who Are English Learners (OPEN ICPSR) Hwang, HyeJin; Duke, Nell	2020-12-20
3.	Exploring the Relationship between Kindergarteners' Buddy Reading and Individual Comprehension of Interactive App Books (OPEN ICPSR) Wang, X. Christine	2019-07-31
4.	Supporting Students Reading Complex Texts: Evidence for Motivational Scaffolding (OPEN ICPSR) Reynolds, Dan	2019-07-13

Studies (2,566) | Variables (2,164) | Series (168) | Data-related Publications (8) | ICPSR Website (10)

Sort by: Relevance

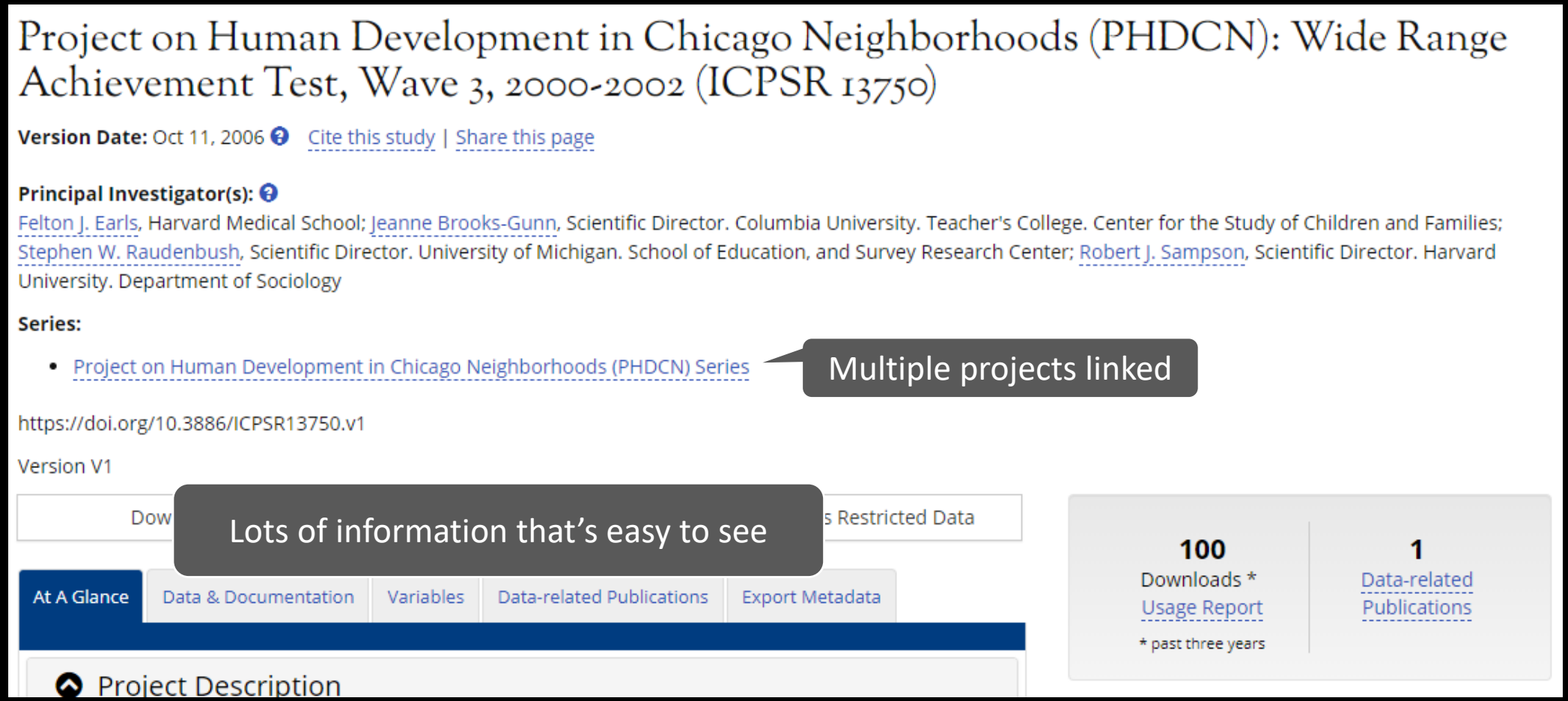
4 5 6 7 8

Select... | Compare

	Var. Name	Label/Question Text	Var. Type	Dataset
☐	RINFRENC	*Q15H. ANSWER QUESTIONS REQUIRING INFERENCE (READING)* How often were the following comprehension topics a primary focus of your instruction in this child's reading/language arts class? Answering questions that require inference. *Taken from: Third Grade Follow-up to the Head Start Impact Study (HSIS), United States, 2007-2008.*	numeric	DS5 - Third Grade Follow-up Teacher's Child Report Data
☐	RAP_PREPCTSTRATEGY3A	*Read Aloud Profile: % read alouds with comprehension strategy3A (new vocabulary) used by adult prior to reading book* *Taken from: Evaluation of Child Care Subsidy Strategies: Massachusetts Family Child Care Study, 2005-2007.*	numeric	DS3 - 2006 One Year Observation Data
☐	RAP_PREPCTSTRATEGY3B	*Read Aloud Profile: % read alouds with comprehension strategy3B (background information) used by adult prior to reading book* *Taken from: Evaluation of Child Care Subsidy Strategies: Massachusetts Family Child Care Study, 2005-2007.*	numeric	DS3 - 2006 One Year Observation Data
☐	RAP_POST_PCT_STRATEG	*Read Aloud Profile: % read alouds with comprehension strategy5 (higher order*	numeric	DS3 - 2006 One Year

Another Project:

- Scrolled down further:



At A Glance | Data & Documentation | Variables | Data-related Publications | Export Metadata

Name	Size	Pre
DS0 Study-Level Files	30 MB	
DS1 Cohort 3	483 KB	
DS2 Cohort 6	495 KB	

R - Home and Life
[PC cohort 00 - 09]

0A.	ALL INFORMATION IN QS. 1 - 9
0B.	TO BE FOUND IN PC ID# |1| | | | | |2

THIS BOX ADDED IN V.2

Now I have some questions about how you spend your time.

1	Do you get a chance to visit with relatives, friends or neighbors in your home or at their home?	Yes.....Continue.....1 No.....Go to Q 2.....2
1A	About how often do you do this? Would you say . . .	Less than once a month.....1 About once a month. .....2 A few times a month.....3 At least a few times a week.....4
2	Do you participate in any child-related organizations, like the YMCA, Boy's and Girl's club, Scouts, or youth programs?	Yes.....1 No.....2
3	Do you participate in any church or religious clubs or activities, not including attending services?	Yes.....1 No.....2

Next I have some questions about family routines, that is whether your family does different things at about

Take Away Points

- Storing Data is not simple.
  - It takes time and money
  - On my first data sharing project, I hired a graduate student
  - Consider budgeting for a data manger on your next grant.

- This is *more* simple if you start early
  - Now, we compile documentation *throughout* implementation
  - Is this project year 1 for you? Write down everything you can!
  - Compile your protocols into one document when you deposit data.

Take Away Points (Continued)

- Make a rule. Write it down. Don’t change it.
  - Applies to protocols, IDs, data collection, data cleaning…
  - This will keep you from needing to fix as many problems later.
  - Make the rules for future you!

- This is not scary
  - Repositories have worked to help make this easy for you
  - You can do it too!

Thank You!

Jessica Logan

Logan.251@osu.edu

Twitter: @JARLogan

1. Data sharing is not simple.
2. It is more simple if you start early.
3. Make a rule. Write it down. Don’t change it.
4. Data sharing is not scary and you can do it!