ICPSR

Managing research and data for reproducibility and transparency

Margaret Levenstein, ICPSR Director

Office of Planning, Research and Evaluation

2019 Open Science Methods Meeting

ICPSR INSTITUTE FOR SOCIAL RESEARCH UNIVERSITY OF MICHIGAN

ICPSR

Founded in 1962 by 22 universities, now consortium of ~800 institutions world-wide

Focus on social and behavioral science data, broadly defined

Current holdings

> 11,000 studies, quarter million files

> 1500 are *restricted studies*, almost always to protect confidentiality

> Bibliography of Data-related Literature with 80,000 citations

Approximately 60,000 active MyData ("shopping cart") accounts

Thematic collections of data about addiction and HIV, aging, arts and culture, child care and early education, criminal justice, demography, health and medical care, and minorities

ICPSR

What is reproducibility?

Can another researcher obtain the same results, using the same data and code?

Can they access the data and code?

If they can, are their results the same?

If not, why not?

Replication versus reproduction

Same substantive inference with other data, specification

ICPSR

Why does reproducibility matter?

1. Knowledge building
   - Is it true?
     - Challenges of p-hacking, especially in a big data world
   - Why is it true?
2. Credibility
   - How do others know it is true?
     - Traditional refereeing process and imprimatur of the academy
       - No longer enough
         - Internet and post-modernism undermined gatekeeper role
     - Confidential and found data confound even the referees

The "crisis of reproducibility" undermines the use of science for evidence-based policy

   Psychology, economics, but also health, others

Sharing is caring

Reproducibility requires sharing data and code
  Respect for study participants
  Minimize burden and increase impact
  Incremental knowledge building
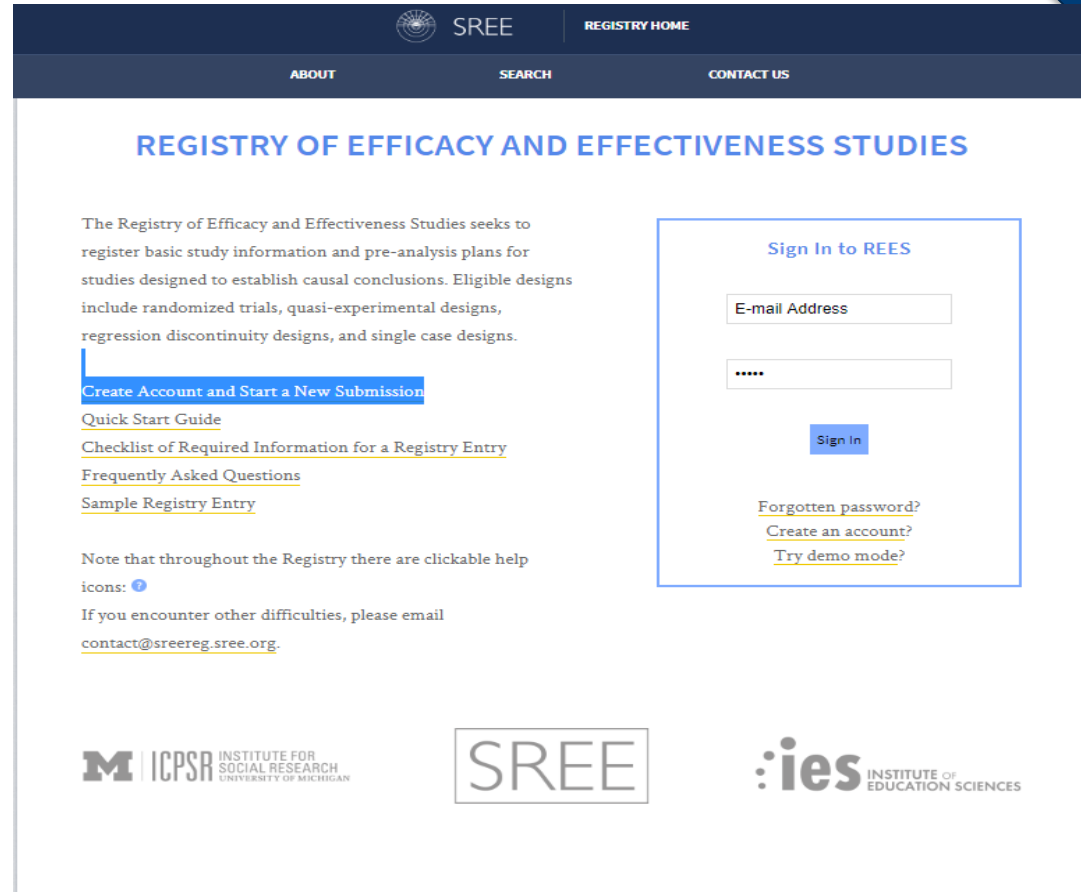 Trust and credibility
Plan for data sharing
  Preregister research
  Data management plan
  Consent statement

ICPSR

Resources for sharing

Preregistration for education effectiveness studies

https://sreereg.icpsr.umich.edu/

Resources for sharing

ICPSR Data Management & Curation

Log In/Create Account

QUALITY   PRESERVATION   ACCESS   CONFIDENTIALITY   CITATION   TOOLS & SERVICES

Additional Resources

- ICPSR's Approach to Confidentiality
- American Statistical Association, Data Access and Personal Privacy: Appropriate Methods of Disclosure Control ↗
- Confidentiality and Data Access Committee (CDAC) forum for staff members of Federal statistical agencies ↗

Recommended Informed Consent Language for Data Sharing

Language to Avoid

Promises in the informed consent can appear to limit an investigator's ability to share data with the research community. In reality, investigators can inform study participants that they are scientists with an obligation to protect confidentiality and still share the study data with the broad scientific community. Many effective means exist to create public-use data files or share restricted-use data files under controlled conditions. That is, data can be modified to reduce the risk of disclosure or shared with additional safeguards while preserving their value for science.

Model Language

Here are two model statements investigators may use in informed consents to describe protection of confidentiality that also allows data sharing.

Sample 1. Study staff will protect your personal information closely so no one will be able to connect your responses and any other information that identifies you. Federal or state laws may require us to show information to university or government officials (or sponsors), who are responsible for monitoring the safety of this study. Directly identifying information (e.g. names, addresses) will be safeguarded and maintained under controlled conditions. You will not be identified in any publication from this study.

Sample 2. The information in this study will be used only for research purposes and in ways that will not reveal who you are. Federal or state laws may require us to show information to university or government officials (or sponsors) who are responsible for monitoring the safety of this study. You will not be identified in any publication from this study.

Known Concerns and Recommended Alternatives

https://www.icpsr.umich.edu/icpsrweb/content/data management/confidentiality/conf-language.html

Consent statements and sharing

Temptation is to promise that no one else will see the data

> Or even that the data will be destroyed
>> This is the direction GDPR has taken

Promise instead to create the most scientific impact while protecting confidentiality

> Separate and encrypt Personally Identifiable Information (PII)

> Restrict use to scientific and evidence-building purposes

> Never reveal information about individual or share with those who try to use to re-identify individuals

Resources for sharing

https://youtu.be/0m5kgYsPwe0

ICPSR

For you old schoolers

ICPSR

Guidelines for Effective Data Management Plans
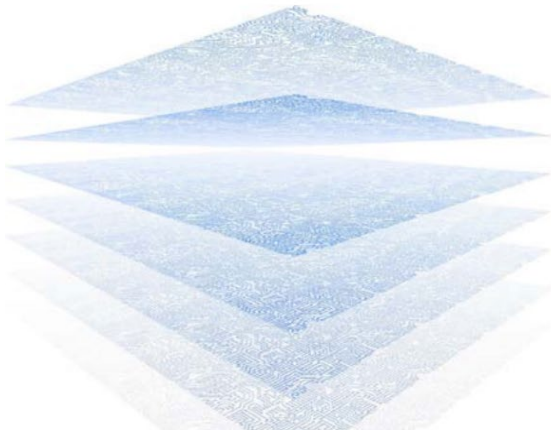
Data Management Plans
Federal funding agencies are increasingly recommending or requiring formal data management plans with all grant applications.

To help researchers meet those requirements, ICPSR offers these guidelines.

Based on our Data Management Plan Web site, this document contains a framework, example data management plans, links to other resources, and a bibliography of related publications.

ICPSR also hosts a blog on data management plans, and a recent webinar on the subject can be viewed on our Web site.

We hope you find this information helpful as you craft a data management plan. Please contact us at netmail@icpsr.umich.edu with any comments or suggestions.

https://www.icpsr.umich.edu/files/datamanagement/DataManagementPlans-All.pdf

ICPSR

Why are DMP important?

Think about data documentation and sharing at the beginning of the project

➢ Improves the research

➢ Makes research reproducible

➢ Reduces cost and increases quality of shared data

Communicates to others

➢ Participants

➢ Funders

➢ Archive

ICPSR

Key elements of DMP

Description of collection (sample, methods)

Short-term storage

Metadata (data about data)

> ➢ Recommendation: standardized, machine actionable

Provenance (especially if you are combining data)

Intellectual property rights

> ➢ Open access means specific licenses

Access policy

Long term preservation

ICPSR

Where to share?

FAIR data

  Findable, Accessible, Interoperable, Reusable

Put your data where it will be

  Found by others

  Preserved in the face of technological change

  Safe for provenance and confidentiality

  Uniquely and persistently identified

  Cited

ICPSR

LINKAGE LIBRARY

Maintaining datasets to support the data linkage community

ICPSR

LINKAGE LIBRARY

Enable researchers to share linked (or linkable) data and linkage strategies

  ➢ Algorithms, code

Compare approaches across projects, datasets, disciplines

  ➢ Improve linkage practices

  ➢ Improve transparency

Build data community

  ➢ Threaded commenting among community members

ICPSR

When to prepare?

Now!

A well-prepared data collection "contains information intended to be complete and self-explanatory" for future users.

ICPSR

Guide to Social Science
Data Preparation
and Archiving

Best Practice Throughout the Data Life Cycle · 5th edition

http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

MANAGING AND
SHARING DATA

UK · DATA
ARCHIVE

BEST PRACTICE FOR RESEARCHERS   MAY 2011

ICPSR

http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Is the data collection complete, accurate, and well-documented?

Documentation

iii

GENERAL SOCIAL SURVEYS, 1972-2010 CUMULATIVE CODEBOOK

(Codebook for the Machine-Readable Data File
General Social Surveys, 1972-2010)

Principal Investigator	Tom W. Smith
Co-Principal Investigator	Peter V. Marsden
Co-Principal Investigator	Michael Hout
Senior Research Scientist	Jibum Kim
Research Assistants	Jaesok Son
	Nicholas R. Nunez
	Matt Gross
	Jerome Gutterman
	Tamila Hill
	Faith R. Laken
	Beatriz Marquez
	Joshua Gagne

NORC Edition

Produced by
National Opinion Research Center
University of Chicago

as part of
The National Data Program for the Social Sciences
2011

This project was supported by
the National Science Foundation
(National Data Program for the Social Science Series, No. 18)

ICPSR

Essential Descriptive Elements

Basic front matter

Variable level details

Methodology

Documentation: Front Matter

Title

GENERAL SOCIAL SURVEYS, 1972-2010 CUMULATIVE CODEBOOK

(Codebook for the Machine-Readable Data File
General Social Surveys, 1972-2010)

Principal Investigator	Tom W. Smith
Co-Principal Investigator	Peter V. Marsden
Co-Principal Investigator	Michael Hout
Senior Research Scientist	Jibum Kim
Research Assistants	Jaesok Son
	Nicholas R. Nunez
	Matt Gross
	Jerome Gutterman
	Tamila Hill
	Faith R. Laken
	Beatriz Marquez
	Joshua Gagne

http://dx.doi.org/10.3886/ICPSR31521.v1

Principal Investigator(s)

ICPSR

Documentation: Front Matter

> **INTRODUCTION**
>
> **DATA COLLECTION DESCRIPTION**
>
> MONITORING THE FUTURE: A CONTINUING STUDY OF AMERICAN YOUTH, 2009 is conducted by the University of Michigan's Institute for Social Research and receives its core funding under grants from the National Institute on Drug Abuse. (The responsible investigators are: Lloyd D. Johnston, principal investigator; Jerald G. Bachman, Patrick M. O'Malley, and John Schulenberg, co-principal investigators.) The research project is unusually comprehensive in several respects: surveys are conducted annually on an ongoing basis; the samples are large and nationally representative; and the subject matter is very broad, encompassing some 1400 variables per year.
>
> The Monitoring the Future Project is designed to explore changes in many important values, behaviors, and lifestyle orientations of contemporary American youth. Two general types of tasks may be distinguished. The first is to provide a systematic and accurate "description" of the youth population of interest in a given year, and to quantify the direction and rate of the changes taking place among them over time. The second task, more analytic than descriptive, involves the "explanation" of the relationships and trends observed to exist.

Description

Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 2009. Johnston, Lloyd D., Jerald G. Bachman, Patrick M. O'Malley, and John E. Schulenberg. Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 2009 [Computer file]. ICPSR28401-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-10-27. doi:10.3886/ICPSR28401.v1

ICPSR

Documentation: Variable-level Details

Frequency	Code	Response	Variable Name	Type/Length
19. On average, what percentage of the students in 1993 who were enrolled in each grade at the beginning of the school year were retained in the same grade (that is, held back) for the next school year? *(For any grade that is not included in your school, circle "N.A.")* (WRITE IN PERCENT)				
a. 7th grade			A19A	num 3
33	0	No 7th grade students retained.		
51		% of students who were retained—range 01-30%		
80	997	legitimate skip/not applicable. School does not have this grade or is an ungraded school.		
8	●	missing		

National Longitudinal Study of Adolescent Health (Add Health), 1994-1995 (National Longitudinal Study of Adolescent Health (Add Health), Wave I School Administrator Codebook.
http://www.cpc.unc.edu/projects/addhealth/codebooks/wave1/index.html

ICPSR

Documentation: Variable-level Details

Variable Name

Frequency	Code	Response	Variable Name	Type/Length
		19. On average, what percentage of the students in 1993 who were enrolled in each grade at the beginning of the school year were retained in the same grade (that is, held back) for the next school year? *(For any grade that is not included in your school, circle "N.A.")* (WRITE IN PERCENT)		
		a. 7th grade	**A19A**	num 3
33	0	No 7th grade students retained.		
51		% of students who were retained—range 01-30%		
80	997	legitimate skip/not applicable. School does not have this grade or is an ungraded school.		
8	●	missing		

ICPSR

Documentation: Variable-level Details

Variable Label

Frequency	Code	Response	Variable Name	Type/ Length
		19. On average, what percentage of the students in 1993 who were enrolled in each grade at the beginning of the school year were retained in the same grade (that is, held back) for the next school year? *(For any grade that is not included in your school, circle "N.A.")* (WRITE IN PERCENT)		
		a. 7th grade	**A19A**	num 3
33	0	No 7th grade students retained.		
51		% of students who were retained—range 01-30%		
80	997	legitimate skip/not applicable. School does not have this grade or is an ungraded school.		
8	●	missing		

ICPSR

Documentation: Variable-level Details

Variable Type

Frequency	Code	Response	Variable Name	Type/ Length
19. On average, what percentage of the students in 1993 who were enrolled in each grade at the beginning of the school year were retained in the same grade (that is, held back) for the next school year? *(For any grade that is not included in your school, circle "N.A.")* (WRITE IN PERCENT)				
a. 7th grade			**A19A**	num 3
33	0	No 7th grade students retained.		
51		% of students who were retained—range 01-30%		
80	997	legitimate skip/not applicable. School does not have this grade or is an ungraded school.		
8	●	missing		

ICPSR

Documentation: Variable-level Details

Question Text

Frequency	Code	Response		Variable Name	Type/ Length
		19. On average, what percentage of the students in 1993 who were enrolled in each grade at the beginning of the school year were retained in the same grade (that is, held back) for the next school year? *(For any grade that is not included in your school, circle "N.A.")* (WRITE IN PERCENT)			
		a. 7th grade		**A19A**	num 3
33	0	No 7th grade students retained.			
51		% of students who were retained—range 01-30%			
80	997	legitimate skip/not applicable. School does not have this grade or is an ungraded school.			
8	●	missing			

ICPSR

Documentation: Variable-level Details

Values

Frequency	Code	Response	Variable Name	Type/ Length
19. On average, what percentage of the students in 1993 who were enrolled in each grade at the beginning of the school year were retained in the same grade (that is, held back) for the next school year? *(For any grade that is not included in your school, circle "N.A.")* (WRITE IN PERCENT)				
a. 7th grade			A19A	num 3
33	0	No 7th grade students retained.		
51		% of students who were retained—range 01-30%		
80	997	legitimate skip/not applicable. School does not have this grade or is an ungraded school.		
8	●	missing		

ICPSR

Documentation: Variable-level Details

Value Labels

Frequency	Code	Response	Variable Name	Type/ Length
19. On average, what percentage of the students in 1993 who were enrolled in each grade at the beginning of the school year were retained in the same grade (that is, held back) for the next school year? *(For any grade that is not included in your school, circle "N.A.")* (WRITE IN PERCENT)				
a. 7th grade			**A19A**	num 3
33	0	No 7th grade students retained.		
51		% of students who were retained—range 01-30%		
80	997	legitimate skip/not applicable. School does not have this grade or is an ungraded school.		
8	●	missing		

ICPSR

Documentation: Variable-level Details

Missing Data

	Frequency	Code	Response	Variable Name	Type/ Length
19.	On average, what percentage of the students in 1993 who were enrolled in each grade at the beginning of the school year were retained in the same grade (that is, held back) for the next school year? *(For any grade that is not included in your school, circle "N.A.")* (WRITE IN PERCENT)				
a. 7th grade				**A19A**	num 3
	33	0	No 7th grade students retained.		
	51		% of students who were retained—range 01-30%		
	80	997	legitimate skip/not applicable. School does not have this grade or is an ungraded school.		
	8	●	missing		

ICPSR

Documentation: Variable-level Details

Summary Statistics

Frequency	Code	Response	Variable Name	Type/ Length
		19. On average, what percentage of the students in 1993 who were enrolled in each grade at the beginning of the school year were retained in the same grade (that is, held back) for the next school year? *(For any grade that is not included in your school, circle "N.A.")* (WRITE IN PERCENT)		
		a. 7th grade	**A19A**	num 3
33	0	No 7th grade students retained.		
51		% of students who were retained—range 01-30%		
80	997	legitimate skip/not applicable. School does not have this grade or is an ungraded school.		
8	●	missing		

ICPSR

Documentation: Variable-level Details

Constructed Variables

15. Siblings

Variable name: **sibname1 - sibname7**

Siblings name. Information about siblings was submitted to the Pension Board when a recruit needed to prove his age in order to receive an age-dependent pension. Sibling names were collected from family Bibles and other sources. If the Pension Board conducted a census search, the generated document also contained siblings= names and ages. Sibling names were also extracted from affidavits and depositions. This variable was cleaned according to the rules for names (see General Information, V.A.2). Comments included the relationship of the sibling to the recruit, especially in the cases when it was a step- or half-sibling, as well as dates and places. SIS and BRO were expanded to SISTER and BROTHER, and 1/2 was changed to HALF.

ILTOT31 – Illegal Activities – Wave 3

The total score was calculated by taking the mean of the z-scores of the following items: ril2ar, ril4ar, ril6ar, ril7ar, ril8ar, ril11ar, ril13ar, ril14ar, ril15ar, ril17ar, ril22ar. Eight of the 11 items need valid responses for a score to be calculated. To address the skewed distribution of the scale, a transformed score was computed by adding 1 to the mean and taking the natural log of that value.

ICPSR

Documentation: Variable-level Details

Notes

```
H00034.00     [H40-SF12-2]                        Survey Year: 2002

             SF12 - ASSESSMENT OF R'S GENERAL HEALTH

In general, would you say your health is ....

NOTE: SF-12(r) Health Survey (Medical Outcomes Trust)
(c) Medical Outcomes Trust and John E. Ware, Jr., All Rights Reserved
SF-12(tm) (QualityMetric, Inc.)

    1232        1 Excellent
    2111        2 Very Good
    1531        3 Good
     563        4 Fair
     145        5 Poor
  -------
    5582

Refusal(-1)           6
Don't Know(-2)        0
TOTAL =========>   5588   VALID SKIP(-4)   7098   NON-INTERVIEW(-5)      0

Lead In: H00033.00[Default]
Default Next Question: H00035.00
```

Skip Patterns

ICPSR

Documentation: Methodology

Sample design: A description of how the cases that appear in the study were selected, including details about target populations, sampling frames, sample sizes, sampling errors, and sampling methods.

Data collection procedures: The methods used to collect the data (e.g., telephone, mail, computer-assisted). Where applicable, this includes the exact instructions and protocols used by interviewers when they collected the data.

Data processing: The activities and quality checks performed on the data collection to generate the final data products from the raw collected data. If files were merged , a full description of the process should be provided.

ICPSR

Documentation: Methodology

<u>Weighting</u>: Where applicable, a description of the criteria for using weights in the analysis of a data collection, including how the weights were created, all weighting formulae or coefficients, a definition of their elements, and an indication of how the formulae are applied to the data.

<u>Confidentiality issues</u>: Where applicable, a discussion of any confidentiality issues in the data, as well as the steps taken to mitigate disclosure risk.

ICPSR

Other Documentation

Questionnaire

User Guide

Handbook

Manual

Report

Table

User Agreement

Errata

ICPSR

Useful Resources: Description

ICPSR, "Guide to Codebooks"

http://www.icpsr.umich.edu/files/deposit/Guide-to-Codebooks_v1.pdf

Institute for Health and Care Research Quality Handbook
http://www.emgo.nl/kc/codebook/

Princeton University Data and Statistical Services, "How to Use a Codebook"
http://dss.princeton.edu/online_help/analysis/codebook.htm

UCLA Social Science Data Archive, "Codebooks"
https://web.archive.org/web/20120601083002/http://dataarchives.ss.ucla.edu/tutor/tutcode.htm

ICPSR

Key Learnings

Ensuring reproducibility will increase the impact of your research

Reproducibility requires sharing data and code

    Where it is preserved and accessible

    Where it is documented and discoverable

Sharing data and code is facilitated by a DMP

ICPSR

Data Jeff wants you to share!