



Trajectories of Secondary Data Use in the Social Sciences

Amy Pienta, Libby Hemphill, Dharma Akmon
ICPSR

Importance of Understanding Data Use

Sponsors

- Accelerate knowledge Discovery
- Increase return on investment
- Greater societal impact

Scientists

- Demonstrate impact (sponsor, employer)
- Maximize use of data
- Contribute to others work

Repositories

- Add relevant data to holdings
- Maximize limited resources for acquisition, curation, preservation
- Improve curatorial decisions
- Ensure FAIR principals

ICPSR Acquisitions

- ICPSR's collection development policy is wide open; serving the needs of the social sciences broadly
- ICPSR is accountable to a large number of stakeholders (Council, ORs, Sponsors, Depositors, Users, Professional Associations)
- Important to archive data that are going to get used
- Curation is costly. Knowing what to curate and what curatorial actions to take would be helpful.

Measuring Data Use

- Measure of researcher/user behavior with data
- Primary use versus secondary use
- Bibliometric Approach – publication counts; citation patterns (e.g. H-Index)
- Repositories: page views, download counts, users; and citation counts.

ICPSR Website in 1997

[ICPSR Home](#) - [Criminal Justice Data](#) - [Aging Data](#)
[contents](#) - [archive](#) - [summer](#) - [contacts](#) - [forums](#) - [sites](#)



Welcome. The Inter-university Consortium for Political and Social Research (ICPSR), located within the [Institute for Social Research](#) at the [University of Michigan](#), is a membership-based, not-for-profit organization serving member colleges and universities in the United States and abroad. ICPSR provides:

Access to the world's largest [archive](#) of computerized social science data.
Training facilities for the [study](#) of quantitative social analysis techniques.
Resources for social scientists using advanced computer technologies.

***** [Table of Contents](#) *****

ICPSR: <http://www.icpsr.umich.edu>
[website technical issues](#)

We wanted to take a long view of data use...but not too long

Access to ICPSR's Data Resources

ICPSR Official Representatives (ORs) at member institutions have access to all the holdings of the Archive free of charge. These data can be provided to researchers and students across the entire campus.

While institutional membership is the main route of access to ICPSR resources, data are also available to individuals at nonmember institutions for a fee.

ICPSR delivers data on a variety of media:

File Transfer Protocol (FTP)

Most of the holdings are available for FTP retrieval. [A list of "FTPable" collections](#) is available; this list changes often as new data are added for FTP. By July 1996, ICPSR's entire collection will be online. [Instructions for using the FTP service](#) are provided.

CD-ROM

ICPSR distributes several CD-ROMS organized around specific substantive topics. Consult this [list of CD-ROMs](#) for more information.

Diskette

Hundreds of collections are available from ICPSR on diskette. A [list of ICPSR data collections on diskette](#) is available.

Magnetic Tape

Currently, ICPSR distributes data written to user specifications on round and cartridge tape. However, as of July 1996, this service will no longer be free. See the new ICPSR [policy on data media and distribution](#) for more information.

ICPSR History

- 2000 – ICPSR began disseminating downloadable public-use data on the web
- 2006 – Data delivery improved at ICPSR so that one can build a dataset containing data usage by study, annually
- We focus on data added in 2006 onward

Our ICPSR Data Set

**N=2,225
Studies
Released
2006+**

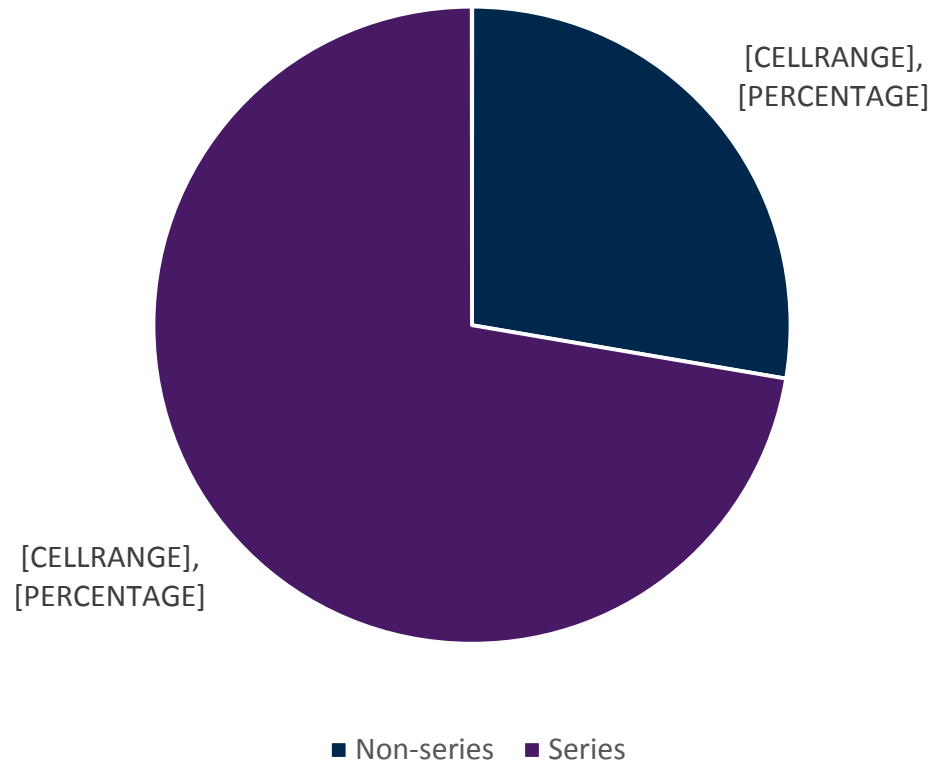
Year of Release	# Studies Released	Percent
2006	173	7.78
2007	202	9.08
2008	270	12.13
2009	210	9.44
2010	196	8.81
2011	217	9.75
2012	177	7.96
2013	275	12.36
2014	224	10.07
2015	281	12.63

Outcome: Data Use

- Unique Users Downloading Files (Data files and/or Documentation files)
- Data Use is measured each Year (2006-2016)

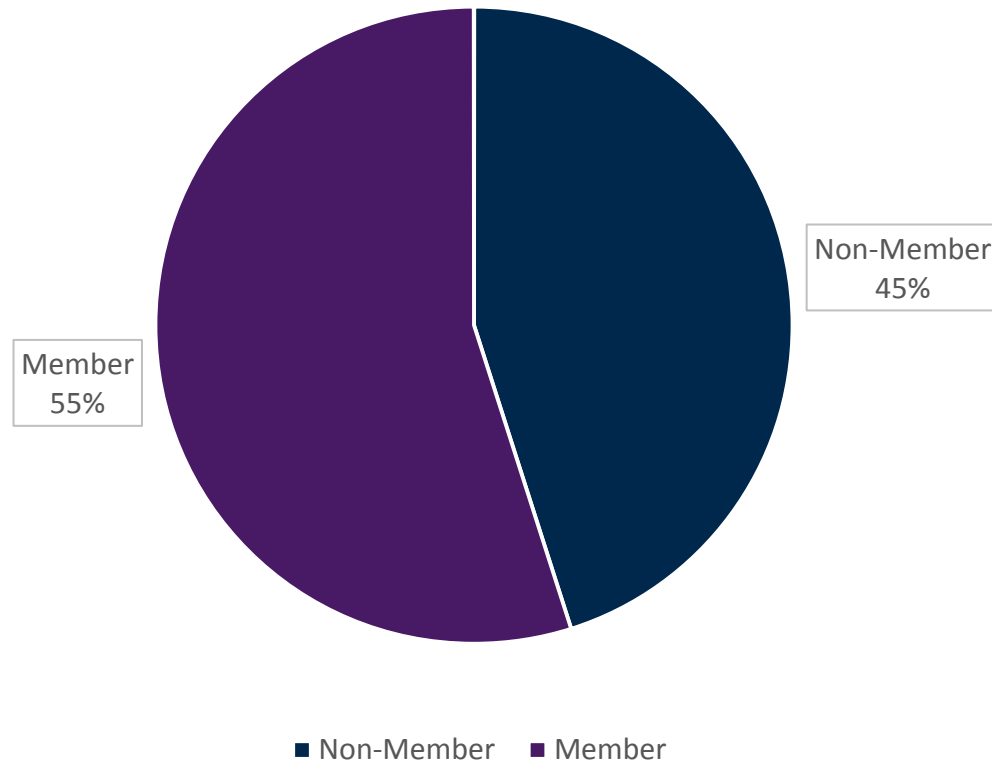
Series Status

Part of ICPSR Serial Collection

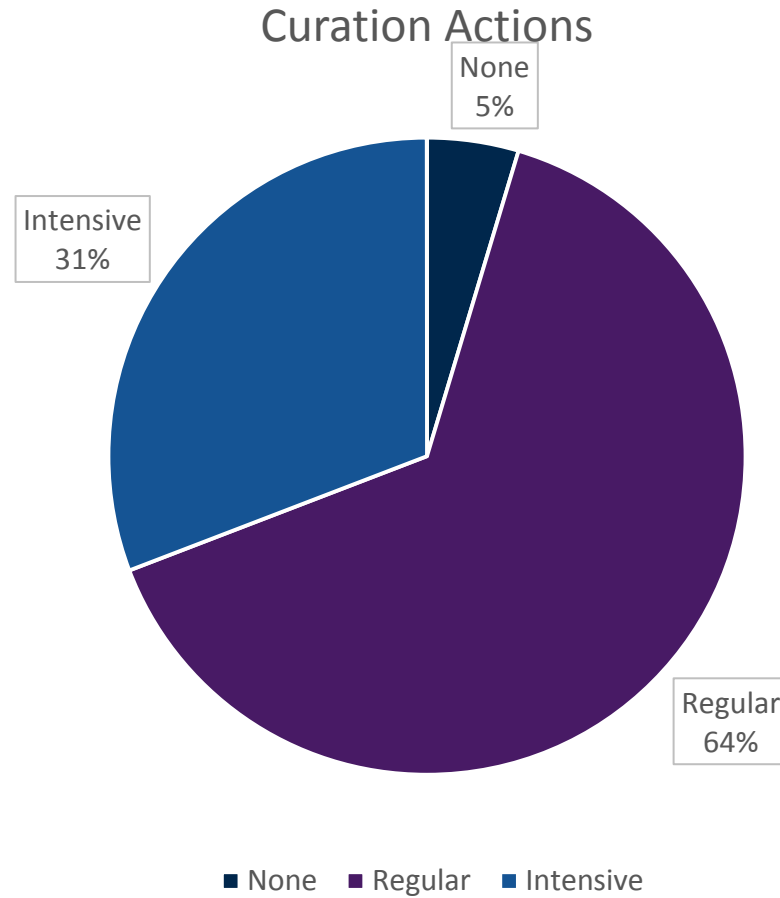


Restricted to Members

Member Access v. Free

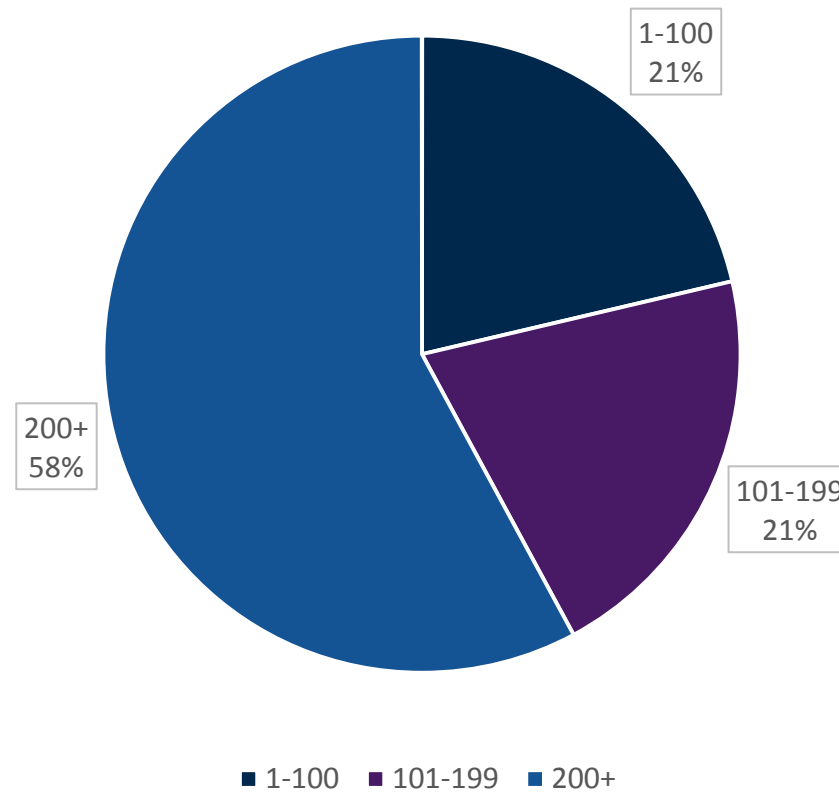


Curatorial Actions



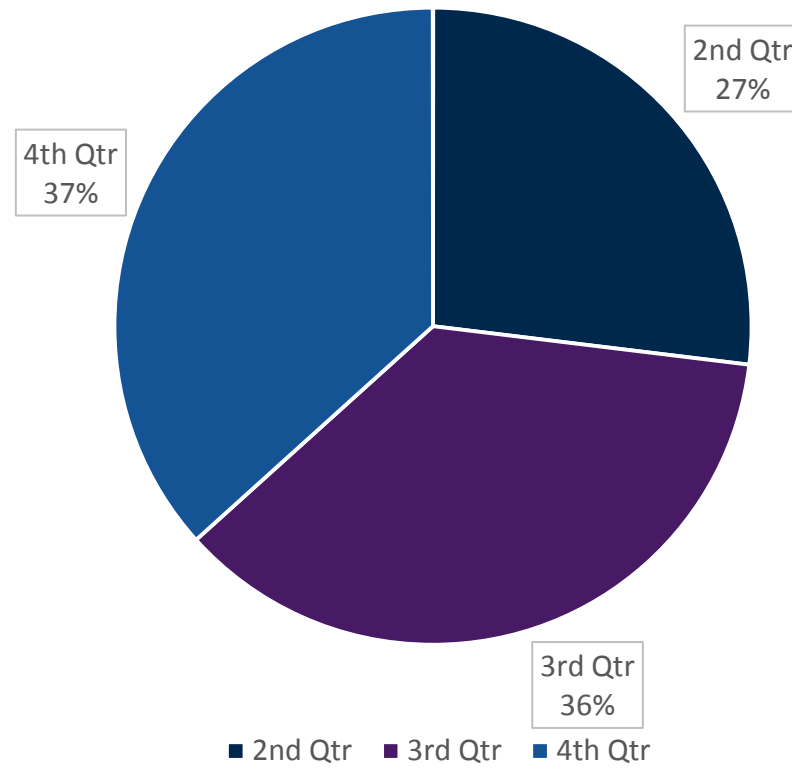
Scope of Data

Number of Variables



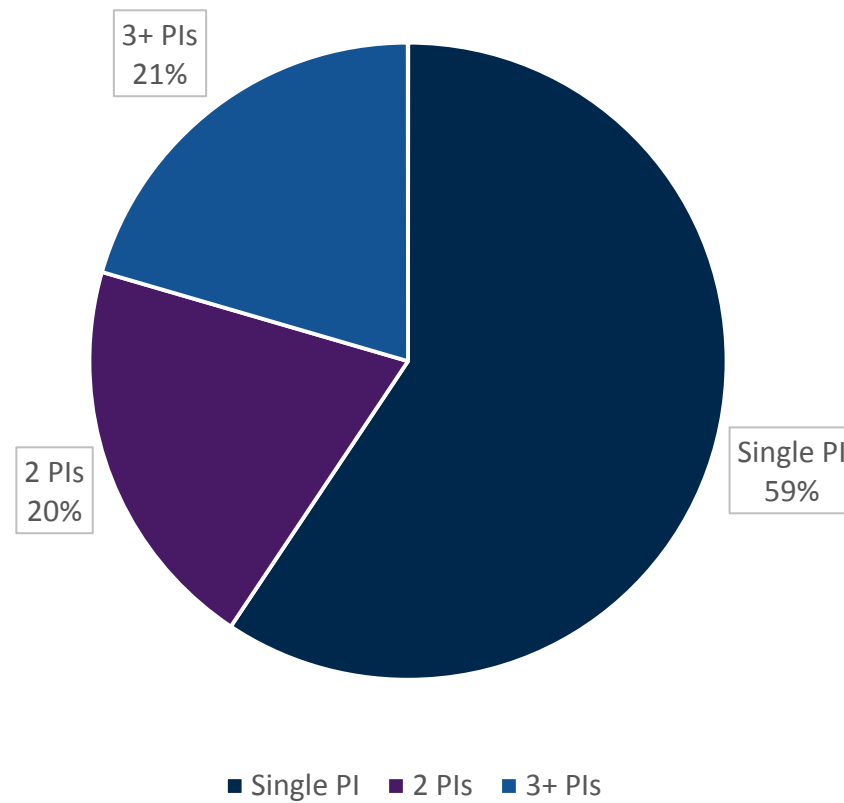
Scope of the Data

Number of Subject Terms



Data Author

Data Authors



Multivariate Approach to Count Data

- GEE – Poisson Regression that offsets the outcome measure by the number of days a study was released and available for use
- Adjusts for the within-study correlation of the annual data points!

Results: Multivariate GEE

	Model 1		Model 2		Model 3	
Intercept	4.941	***	5.025	***	4.161	***
Series=1	-.544	***	-.517	***	-.547	***
Member (v. free)			-.779	***	-1.005	***
Curation Actions						
None (replication)			-1.614	***	-1.359	***
Regular			Ref.		-	
Intensive (SDA)			.622		.584	
200+ Vars (v. 1-199)					.588	***
# Subject Terms					.029	***
# PIs						
1 PI					Ref.	ns
2 PIs					.082	
2+ PIs					.097	

Conclusions

- Studies that are wider in scope get more use.
- Self-published studies and/or replication data are rarely used, compared to ICPSR's curated collections.
- Series are highly used, but a new data release in a series dampens use of earlier studies. Users want the newest data

Limitations

- We are really interested in curatorial actions that affect user behavior
 - Selection v. Causation???
 - ICPSR may be good at predicting data use and takes appropriate curatorial actions on a study
- We investigated data use overall, and not yet trajectories of use. That is next!
- Other measures:
 - Other data use measures
 - Citation counts (volume of primary use before release)



Thank you!

Amy Pienta

apienta@umich.edu

734-615-7957



@apienta