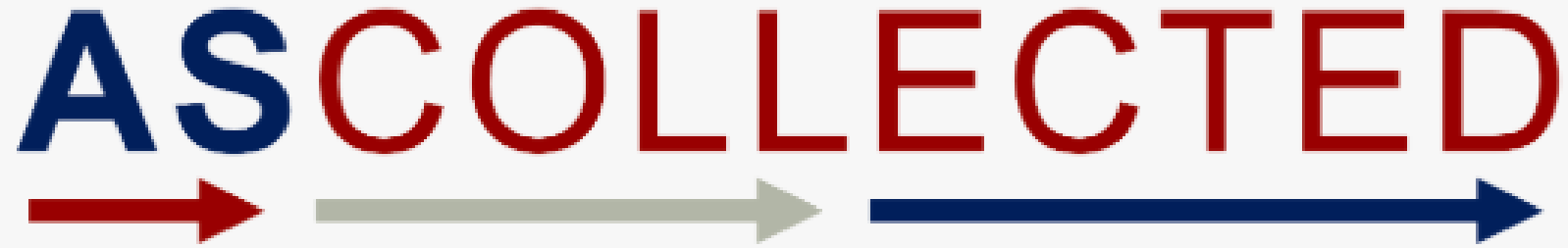


ASCOLLECTED

The logo for ASCOLLECTED features the word in a bold, sans-serif font. The letters 'A' and 'S' are dark blue, while 'COLLECTED' is red. Below the text are three horizontal arrows: a red arrow under 'A', a grey arrow under 'COLLECTED', and a dark blue arrow under 'ED'.

→ → →

The Only Platform for Documenting Results Provenance in Science

Uri Simonsohn

Prof. Behavioral Science
Esade Business School, Barcelona

The Wharton Credibility Lab

- 2015: AsPredicted Pre-Registration Made Easy
- 2020: ResearchBox Posting Data/Code/Materials Made Easy
- 2025: AsCollected Results Provenance (Easy from the start)

Pre-Registration vs Results Provenance

AsPredicted vs AsCollected



Pre-Registration
What's actually in the study?

Results provenance
How, when, where **results** made



NF-FUSS-568207B/7F LAP 30047884

Tells us: production plant, date, & batch

Goals

Pre-registration:

Main: Separate exploratory from confirmatory

Incidental: thinking ahead

Results Provenance:

Main 1: Prevent data & results contamination (e.g., fraud & innocent errors)

Main 2: Facilitate cleanup when contamination is detected

(just like [food traceability](#))

Incidental 1: Culture of code-review

Incidental 2: Credible credit to research doers

What exactly is “results-provenance”

1. When were data collected
2. Where
3. How
4. Who cleaned them
5. Who analyzed them
6. Who inspected the code
7. Who has copies of code & data

Still very much a **trust** system

But now we know **who** we are trusting and with **what**.

(and we know how to **verify** if needed)

CredibilityLab Key Design principles

- Simplicity for
 - Authors
 - Readers
- Standardization
- Very much ***not***
 - “dump your 60 page Word document here...
 - ... lets hope it contains what it should...
 - ...and hope readers can find it...

Next. Credibility Lab principles in AsCollected

1. For Readers

2. For Authors

For Readers: every AsCollected is two tables

- **Table 1:** Who
- **Table 2:** When, Where, How
- Both in one URL like https://ascollected.org/ABC_DEF



- **Next.** Example of Table 1

AsCollected #2 - 'Time Pressure and Moral Dilemmas'

https://ascollected.org/ZJ6_Z47

Authors (anonymous for now)

1. Author 1
2. Author 2
3. Author 3

Dates

Made Public: 2025-10-27

Latest update: 2025-10-27

Show introduction

- **Next.** Table 2.

Table 2: Data collected

	When	Where	Data ID	Non-authors	Additional Info
Study 1	1 Obtained 2024-07-01	1 Obtained from https://moralchoice.mbc.com/archive	1 Unique data ID? No	1	1 Anything else: The data are freely available from that website, and have been archived in the webarchive. Obtaining the data: Full-length recordings of the game show were downloaded from the above URL. Four research assistants [anonymized] watched and coded each episode, putting their scores on a shared google sheet. Data cleaned: With code

Notes:

Numbers on top-right of cell represent the author(s) providing the information.

Next. Credibility Lab principles in AsCollected

1. For Readers

2. For Authors

New Project

 Your Projects

WHAT IS ASCOLLECTED?

A platform for documenting results-provenance in scientific publications; when and how data were obtained and who cleaned & analyzed them. Information for an entire paper is presented in two standardized tables.

Ex 1 (experimental) [AsCollected.org/ZJ6_Z47](https://ascollcted.org/ZJ6_Z47)

Ex 2 (observational) [AsCollected.org/G2Q_W7V](https://ascollcted.org/G2Q_W7V)

HOW DOES IT WORK?


- Information is provided answering short form
- Separately for each data source
- Authors may add/edit information at any time
- For peer review, information is anonymized
- Edits post-publication have track-changes on

WHAT IS IT FOR?

- Hopefully will induce norm of documenting results-provenance, making scientific fraud very difficult
- Disclosing roles in data analysis
 - Gives *credible* credit where it is due
 - May induce norm of code-review in science
- Transparency helps detect honest errors

What research design best fits the study


If more than one, you can later add different designs

 **Lab data**

A study conducted in a physical space specifically set up for research, either permanently (e.g., a behavioral lab) or temporarily (e.g., a classroom). Data may be collected on-site or via a web interface while participants are in that location.

 **Online study**

A study conducted remotely over the internet (e.g., Qualtrics survey, Prolific participants).

 **Field experiment/data collected by authors**

Original data collected by authors (e.g., an A/B test, a policy experiment, a survey of employees, interviews, qualitative observational research, data scraped from a website, etc.)

 **Private data**

Observational data obtained from a third party, not readily available to others (e.g., data purchased from a data provider, data obtained from a government agency after an approval process or fee, data obtained with an NDA, data provided by a friend of the authors who owns a small business)

 **Public data**

Data which are publicly (freely) available in the format they were obtained (e.g., replication package of previously published paper, public census data, PSID, World Values Survey, etc.)

 **Just code**

Simulated data, or standalone code that uses a datasource described elsewhere in the project (e.g., to document code-provenance separately for different analyses of the same data)

Next

Edits: All answers can be edited at any time by any author

Multiple datasources in one study? [see info](#)

Datasource name

[delete datasource](#)

Tip: Give a name that describes the data. You currently have 1 other datasources in this study:

- 1. Private data [✎](#)

Questions about data provenance

Tip: You can always skip questions and answer them later

How were data collected?

- online
- lab computer saving locally
- paper & pencil
- experimenter took notes
- external device (e.g., heart monitor, eye-tracking)
- other

What's the physical location of the lab?

e.g., computer lab in [Smith Hall], at [The Ohio State University]

Information you put inside [squared brackets] will be shown as "[anonymized]" until you deanonimize the project

Date when first observation was collected

If you don't have the exact date use xx, e.g., "2023-05-xx"

Date when last observation was collected

If you don't have the exact date use xx, "2023-05-xx"

Lab Study ID

e.g., study or session number at the lab where study was run; enter "none" if not applicable.

Name research assistants (RAs) and/or lab manager who actively participated in data collection during the lab study.

e.g. We received assistance from two RAs [Ludovic Smith & Jennifer Sepulveda] and the lab manager [Benjamin Nguyen]

What research design best fits the study

If more than one, you can later add different designs

 **Lab data**


A study conducted in a physical space specifically set up for research, either permanently (e.g., a behavioral lab) or temporarily (e.g., a classroom). Data may be collected on-site or via a web interface while participants are in that location.

 **Online study**

A study conducted remotely over the internet (e.g., Qualtrics survey, Prolific participants).

 **Field experiment/data collected by authors**

Original data collected by authors (e.g., an A/B test, a policy experiment, a survey of employees, interviews, qualitative observational research, data scraped from a website, etc.)

 **Private data**

Observational data obtained from a third party, not readily available to others (e.g., data purchased from a data provider, data obtained from a government agency after an approval process or fee, data obtained with an NDA, data provided by a friend of the authors who owns a small business)

 **Public data**

Data which are publicly (freely) available in the format they were obtained (e.g., replication package of previously published paper, public census data, PSID, World Values Survey, etc.)

 **Just code**

Simulated data, or standalone code that uses a datasource described elsewhere in the project (e.g., to document code-provenance separately for different analyses of the same data)

Next

Briefly describe the data (this is especially important if this study has more than one datasource)

Keep in mind readers will read full study description in your paper, so be brief: e.g., 'quantity and price of pastries sold'.

Information you put inside [squared brackets] will be shown as "[anonymized]" until you deanonymize the project

How generally accessible are these data directly from the 3rd party?

- *Accessible with restrictions: anyone passing a screening and/or paying fee could obtain from 3rd party (e.g., Geodata in the PSID, Comscore, etc.)*
- *Exclusive access: others authors would have substantial difficulty (e.g., obtained through personal connection or through lengthy NDA)*

- Accessible with restrictions
 Exclusive access

Name the third party (with a brief description when they are less well-known)

(e.g., "Joe's Shoes, a store in Wynwood Mall in London, UK", "Vanguard, the investment firm", "Amtrak", "UNICEF")

Are you allowed to name the 3rd party providing the data?

- Yes
 No

You may enter here any additional information.

(e.g., details on agreement not to share data, additional relevant dates, etc). If nothing else, write 'No'.

Nothing else.

Information you put inside [squared brackets] will be shown as "[anonymized]" until you deanonymize the project

If you cannot name source because of NDA

Do you have a contract or explicit agreement prohibiting sharing the identity of the 3rd party?

- Yes
 No

Do all authors of this paper know the identity of the data provider (both the organization and the name of the contact providing the data)

- Yes
 No

Outside of the author team, who else formally or informally partook in the agreement with the third party to keep them anonymous? (answer "only authors" if nobody else did)

(e.g., Department chair [Jon Velodromo (Yale)], Office of General Counsel [Victoria Prescott (Wharton)], Department business administrator [Melissa Black (OSU)])

Information you put inside [squared brackets] will be shown as "[anonymized]" until you deanonymize the project

And if you don't have an NDA

Do you have a contract or explicit agreement prohibiting sharing the identity of the 3rd party?

- Yes
- No

Do all authors of this paper know the identity of the data provider (both the organization and the name of the contact providing the data)

- Yes
- No

Explain why, in the absence of a contract, you are not at liberty to indicate identity of the data provider.

Information you put inside [squared brackets] will be shown as "[anonymized]" until you deanonymize the project

Would documenting results provenance prevent fraud?

As the team with the most experience with fraud in soc science...

... we think so.

Let's concretely consider a few cases.

Concrete Cases

1. Lacour (UCLA->Princeton | door-to-door canvassing)
2. Toner-Rodgers (Star MIT PhD student | AI material science)
3. Sanna (Michigan | study in auditorium)
4. Ariely (possibly most famous social scientist |
“don’t know who sent/received insurance data”)
5. Gino (HBS, altering spreadsheets)

Fraud's most direct victims are coauthors

They are also science's first line of defense

Victims/defenders

- Con men (and women) create and exploit ambiguity
- Uncomfortable for coauthors to disambiguate
- AsCollected gives excuse to be an A**hole and ask questions
- AsCollected also protects you from bad data you never touch
- Imagine Gino were innocent, and she was:
 - RA, or maybe a
 - Gift card scammer, or maybe an
 - Envious BU prof, or maybe a
- Set her up
- ...she'd value having used AsCollected at >\$25 million
- **You** can use it for free!



Traceability:

Prevention

Cleanup

ManyCoauthors.org

(precursor to AsCollected)

- After Gino case
- 135 co-authors asked about all of Gino's papers
 - Did she collect the data?
 - Is the data posted?
- A baby (retroactive) AsCollected

Example 1:

... Filter by anything, e.g., JPSP, Smith, authenticity



Clear

[show](#) non-empirical papers

indicates information has been provided by authors but none of them have yet made it public

ID# ↕	Title (click paper title for more details to get paper)↕	Journal ↕	Year ↕	Studies Total ↕	Studies by Gino ↕	Authors asked ↕	Authors responded ↕	Living document ↕	Posted Data ↕	Latest update ↕
1	Different Roots, Different Fruits : Gender-based Differences In Cultur [...] Tost, LP; Hardin, AE; Roberson, JW; Gino, F	A of M	2022	5	0	3	3		✓	2023-10-29

?



Many Co-Authors

ID#	Title (click paper title for more details to get paper)↕	Journal ↕	Year ↕	Studies Total ↕	Studies by Gino ↕	Authors asked ↕	Authors responded	Living ↕ document	Posted Data ↕	Latest update ↕
62	Compared To Men, Women View Professional Advancement As Equally Attainable [...] Gino, F; Wilmoth, CA; Brooks, AW	PNAS	2015	9	0 Don't Know: 9	2	2	✓ (72 words)		2023-11-06
63	Hormones And Ethics: Understanding The Biological Basis Of Unethical C... [...] Lee, JJ; Gino, F; Jin, ES; Rice, LK; Josephs, RA	JEP:G	2015	2	0	3	2	✓ (24 words)		2023-11-03
64	Reducing Bounded Ethicality: How To Help Individuals Notice And Avoid ... [...] Zhang, T; Fletcher, PO; Gino, F; Bazerman, MH	Org Dynamics	2015	1	0	3	2	✓ (96 words)	✓	2023-10-23
65	The Slow Decay And Quick Revival Of Self-deception Chance, Z; Gino, F; Norton, MI; Ariely, D	Frontiers	2015	2	0	3	3			2023-11-04
66	Prosocial Norms In The Classroom: The Role Of Self-regulation In Follo... [...] Blake, PR; Piovesan, M; Montinari, N; Warneken, F; Gino, F	JEBO	2015	1	0	4	2	✓ (182 words)		2023-10-26
67	The Moral Virtue Of Authenticity: How Inauthenticity Produces Feelings ... [...] Gino, F; Kouchaki, M; Galinsky, AD	Psych Science	2015	5	5	2	2	✓ (1030 words)	✓	2023-11-06
68	Smart People Ask For (my) Advice: Seeking Advice Boosts Perceptions Of ... [...] Brooks, AW; Gino, F; Schweitzer, ME	Mgmt Science	2015	8	1 Don't Know: 7	2	2			2023-10-23

WTAF?

Back to the present: AsCollected

- **Using AsCollected**
 - Makes life of fraudsters more stressful
 - Makes life of potential victims less stressful
- **You probably don't think you will ever coauthor with a fraudster**
 - Nobody does
 - Including the 135 Gino coauthors
- **But even if you are right,**
 - Using AsCollected is still a public good.
 - Think of the creeps.

- Onboarding at Wharton: We all keep doors open
- So that creeps keep theirs open
 - (and we are not wrongly suspected of being creeps)



- If **we all** use AsCollected
- Fraudsters **have to** use AsCollected

Roll out strategy

- This talk. You can start this afternoon
- Management Science: required for all submissions
- Psychological Science: submissions not sharing data
- Public announcement when Mgmt Science implements