2SESSION LANDING PAGE

[HACKATHON]

Making open psychological datasets more accessible and useful for research and teaching

Session leaders: Cameron Brick
Contact information: brickc@gmail.com
Location: Willem-Alexander Zone 3 in Engels
Day/time: Sunday 13.45 - 15.15 (& upcoming session Tuesday)

Abstract: Secondary analysis of large, open datasets can improve replicability and reproducibility over small new studies, e.g., through accessibility, high power, generalizability, and validated measures. Between-subjects moderation designs particularly lack power. However, these resources are underused in research and education. Student research often has low inference quality, and educators may lack access to use these tools for pedagogy. We have an index of 80 datasets (mostly large, representative surveys with many non-psychological variables) and some meta-data (constructs, years, location, etc.). Attendees will plan and carry out improvements of the index and its accessibility. Groups may: develop a metadata structure; create metadata; build handouts to help teachers integrate activities and assessments; develop resources to help students at all levels do primary exploratory or confirmatory research; plan strategies to share the resources; or make other contributions. tinyurl.com/sips2019brick https://osf.io/dh25g/

Pre-requisites: Laptops recommended.

Slack Channel: **https://sips-2019.slack.com/messages/h_data_accessibility**
OSF Project Page: https://osf.io/ag23c/

Working documents:

Initial spreadsheet of datasets (google drive)

(see next page)

Session plan (Tues hackathon)

Introduce topic & clarify scope (type of dataset)
Depending on how many attend, briefly summarise interest/abilities
Work group suggestions & discussion
Divide into groups and work. Cameron will float around to support.
Regroup and summaries from each group. Any future plans.

Current action points

1. ONGOING - *Expanding the list of which datasets exist*: especially link codebook(s)
    a. Adding key metadata columns & content
2. DONE - Solicited existing solutions.
    a. https://rachelrenbarger.wordpress.com/2018/03/13/how-to-find-public-use-datasets-for-your-research/
    b. https://tobiasdienlin.com/2019/03/03/publicly-available-open-datasets/
3. DONE - Email database controllers (partial list, anyway
4. **NEEDS HELPER -** Go through here and integrate data sources from this other list: http://bit.ly/2NRok9k (rows 1-19 finished. Can you help with 20-onward?)

Need project leader

5. How to manage the difficulty identifying if someone else has published a certain finding from open data? how to search effectively for particular published effects
6. Develop tips with for working with large-scale data: codebook problems, measurement error problems (eg. brief personality scales), duplicate work problem with data cleaning
7. Data cleaning: identify how do people do it, share, problems
8. Adopting or developing guides or resources to encourage use in research and teaching, e.g., how to supervise BA or MA psychology thesis projects based on existing data. Resources:
    a. https://forrt.netlify.com/
    b. https://psyteachr.github.io/
9. Adopt a machine-readable standard for better sharing, stability, and usability? https://github.com/psych-ds/psych-DS See later session by Melissa Kline: Less data cleaning, more data adventures with Psych-DS and also Ruben Arslan: Document Your Data hackathon
10. (add anything)

Email sent to: Julia Rohrer, Emorie Beck, Markus Jokela, Rich Lucas

Dear researcher,

Psychologists are not using large-scale, secondary, free data to its potential. Major problems are not knowing what exists, what it contains, or how to use it without a huge investment of coding. After our SIPS hackathon, we now have a list of 80 free, large datasets with psychology data. **YOU:** Do you have codebooks, cleaning codes, or summaries of such data sets (e.g, of the World Value Survey)? Please:

1. Email us in any format: how do you organize these datasets, have you summarized their contents, do you have a resource of many codebooks, and generally how are you solving this problem? Let's reduce duplicated work.
2. Contribute to the Google Sheet of existing data (add datasets and/or metadata like years, topic).
3. Any other input/contributions welcome. Feel free to forward.

Emails to Cameron Brick, brickc@gmail.com

More info

In the workshop on making existing, large-scale psychological datasets more accessible and useful we identified two key barriers for re-use:

1. **Lacking datasets and metadata** with general description of dataset size, structure, and contents interesting to psychologists. This is currently a Google sheet open to editing by all.

2. **Lacking code for cleaning**, because many valuable datasets have documentation that is spread-out, requires tacit knowledge (or emails to the data managers) to work with, and also because data cleaning is a lot of work and a common source of errors. This might take the form of a collection of Github code repositories.

Best,

[see list of contributors]

(Request from researchers using your dataset at the University of Cambridge)

Dear researcher or data provider,

We are a collaborative group from the Society for the Improvement of Psychological Science (SIPS) working on a project to **make secondary data more visible**.

Researchers are not using large-scale, secondary data to its potential. We often do not know what datasets exist, what they contain, or how to use to access them. A partial solution would be a simple list of datasets, aggregators, and basic metadata. Other curated lists exist, like at ICPSR, but there is a complexity/useability trade-off and we are building a accessible list.

Do you have five minutes to add the basic description of your dataset or resource? Please fill out this Google Sheet. We are looking for information like:

Years; Total number of participants; Access restriction (e.g., application only; any fees; other restrictions); Link(s) to codebook or search engine; Topics (e.g., household; physiological; well-being)

Alternatively, you can send us a short email with a summary of this information, including the themes covered by your dataset.

We are very grateful for your time and help! Thank you.

[see list of contributors]