An applied introduction to good practices in preregistration

Antonio Schettino, Ph.D.
Senior Advisor Open Science
Erasmus Research Services
schettino@eur.nl

Caspar van Lissa, Ph.D.
Associate Professor Social Data Science
Open Science Community chair
Tilburg University
c.j.vanlissa@tilburguniversity.edu

Erasmus University Rotterdam

slides: *https://osf.io/tfnkc/*

OUTLINE (WORKSHOP)

- What is preregistration?
- How to preregister on the OSF?
- Preregistration As Code (PAC)

Antonio Schettino

OUTLINE (FIRST PART)

- What is preregistration?
- Why preregistration?
- OSF: Hands-on session
  - Create a preregistration
  - Browse examples of public preregistrations

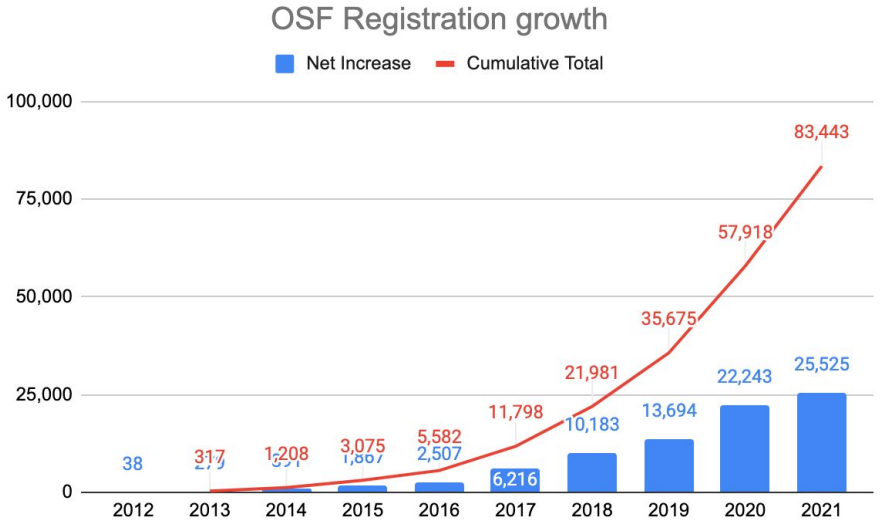
OUTLINE (SECOND PART)

- What is Preregistration-As-Code (PAC)?
- An applied example of PAC from ManyAnalysts
- An ideal example you can build on for your work

What Is Preregistration?

A document that:

- specifies research design and analysis plan *before* the project starts
- is uploaded on dedicated repositories
- is made public, either immediately or after an embargo period



Registration as resource type in Datacite schema (OSF, 2022)

A Community-Sourced Glossary of Open Scholarship Terms (Parsons et al., 2022)

Antonio Schettino

Types of preregistration

Common elements

- research question
- methods
- planned analysis

Unreviewed

- upload on public repository
- collect & analyze data
- write paper
- publish (**?**)

Reviewed (Registered Report)

- submit to participating journal (*list*)
- protocol is peer-reviewed
- in-principle acceptance (IPA)
- collect & analyze data
- update protocol with results & discussion
- publish (**!**)

The past, present and future of Registered Reports (Chambers & Tzavella, 2022)

Antonio Schettino

Is preregistration *necessary*?

NO, if the goal is:

- to test vague theories
  - Good theories are **hard to vary**. They:
    - explain what they are supposed to explain
    - are consistent with other good theories
    - cannot easily be adapted to explain anything
- Vague theories can be adapted **regardless of preregistration**
  - Example: professor-priming effect (*Dijksterhuis & van Knippenberg, 1998*)
  - Registered Replication Report (*O'Donnell et al., 2018*) does not confirm original findings
  - Flexible theory → new moderators (e.g., sex differences, awareness; *Dijksterhuis, 2018*)

Arrested Theory Development: The Misguided Distinction Between Exploratory and Confirmatory Research (Szollosi & Donkin, 2021)

Antonio Schettino

Is preregistration *necessary*?

NO, if the goal is:

- to distinguish between **confirmatory** and **exploratory** analyses
  - this distinction is unclear and/or irrelevant. Example:
    - Researchers preregister study hypothesizing that A → B
    - During data analysis, they discover a paper in the literature claiming that A → C
    - Does the analysis A → C qualify as confirmatory or exploratory?
  - exploratory research is not bad!
    - *"Exploratory research [...] is not synonymous with serendipity but [...] a deliberate and systematic attempt at discovering generalizations that help us describe and understand an area about which we have little or no knowledge [...] it is analogous to topographically mapping an unknown geographical region."* (*Devezer et al., 2021, p.19*)

Does preregistration improve the credibility of research findings? (Rubin, 2020)

Antonio Schettino

Is preregistration *necessary*?

NO, if the goal is:

- to prevent *HARKing* (**H**ypothesizing **A**fter the **R**esults are **K**nown)
  - circular reasoning can be spotted without knowing ***when*** hypotheses were generated
  - Three types of HARKing:
    - *CHARK*ing (**C**onstructing **H**ypotheses **A**fter the **R**esults are **K**nown): no independence from observed evidence (overfitting), should always be disclosed
    - *RHARK*ing (**R**etrieving **H**ypotheses **A**fter the **R**esults are **K**nown): independent from observed evidence, can predict and be falsified by observed evidence (ethical to disclose original source)
    - *SHARK*ing (**S**uppressing **H**ypotheses **A**fter the **R**esults are **K**nown): can artificially inflate the perceived veracity of published conclusions, should always be disclosed unless suppressed hypotheses are unrelated to final conclusions or based on non-severe tests

When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress (Rubin, 2017)

Is preregistration *necessary*?

NO, if the goal is:

- to prevent *p*-hacking: run many statistical tests, report only significant ones (*Simmons et al., 2011*)
  - full disclosure of data collection stopping rule, data exclusions, measures, and manipulations
  - logical and principled justifications for non-standard data exclusions and analytical approaches
  - public access to data analysis procedures
  - public access to research materials, data, and coding information
  - report results of robustness analyses

Does preregistration improve the credibility of research findings? (Rubin, 2020)

Antonio Schettino

Is preregistration *necessary*?

NO, if the goal is:

- to reduce *researcher degrees of freedom, rDF* (*Wicherts et al., 2016*)
  - Arbitrary choices in study design, data collection, analysis, and reporting
  - Can be (ab)used, e.g., for *p*-hacking
  - Also occur naturally, even in preregistered studies
    - See Many Analysts project about religiosity and well-being (MARP: *Hoogeveen et al., 2022*)
    - Preregistration-As-Code to mitigate rDF (*Van Lissa, 2022*)
  - Effect of rDF can be studied using multiverse analysis (e.g., *Young et al., 2022*)

Does preregistration improve the credibility of research findings? (Rubin, 2020)

Antonio Schettino

Is preregistration *necessary*?

NO, if:

- Researchers are testing **hard-to-vary theories,** producing **falsifiable hypotheses**
- Researchers are transparently reporting everything they do
- (Public) Availability of **data**, **code**, and **materials**
- Researcher degrees of freedom are mitigated

Are these assumptions plausible in your research field?

Antonio Schettino

Is preregistration *useful*?

YES:

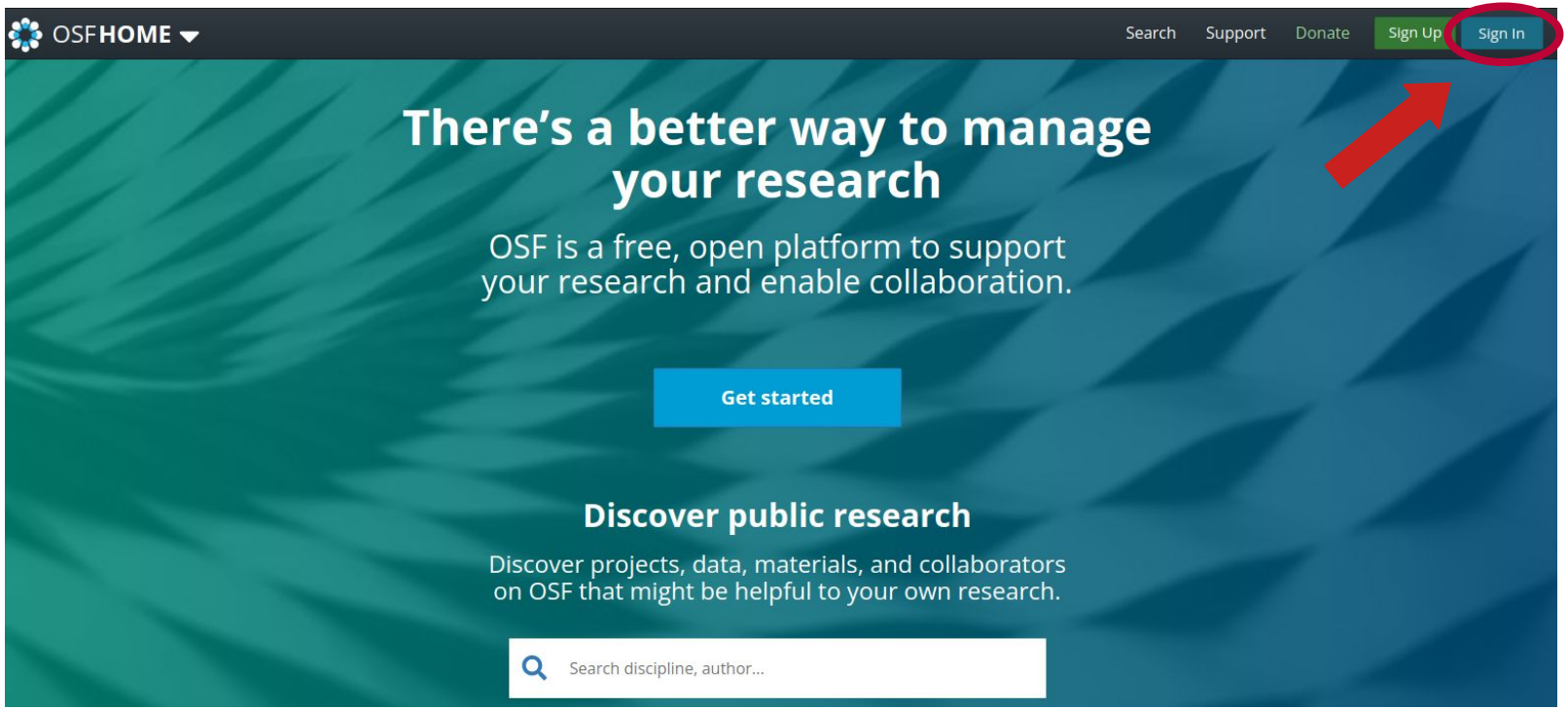
- **Calibrate** confidence on reported results and interpretations
  - especially in absence of other open research practices (e.g., open data, open code, …)
- **Transparent** communication to (possibly) highlight cognitive biases, e.g.:
  - *Confirmation bias*: prefer information that supports our beliefs
  - *Hindsight bias*: “I knew it all along”
- **Improve** study design
  - think more carefully about research plan
  - solicit and incorporate peers’ feedback when most valuable (i.e., before starting!)

Making ERP research more transparent: Guidelines for preregistration (Paul et al., 2021)

HANDS-ON SESSION

Log In

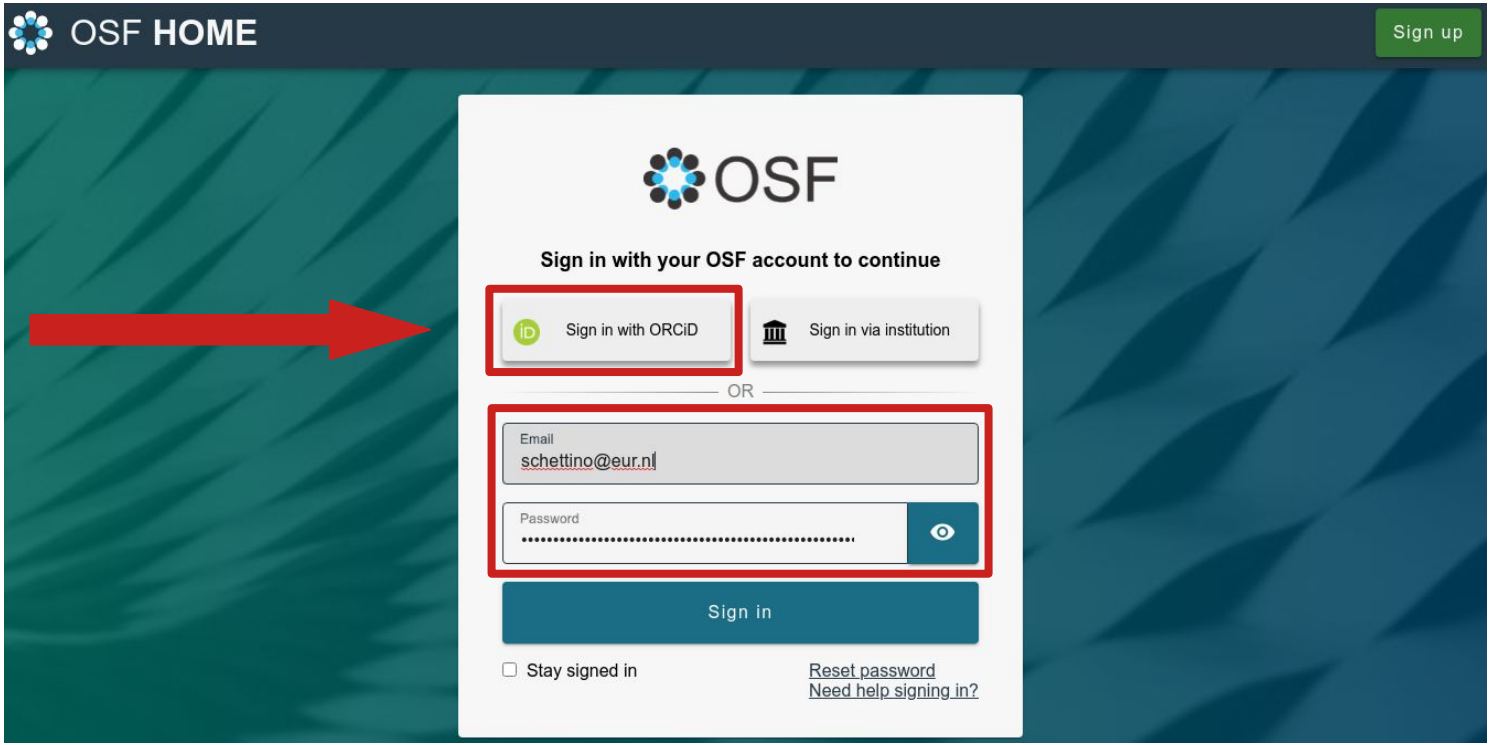
https://osf.io/



Antonio Schettino

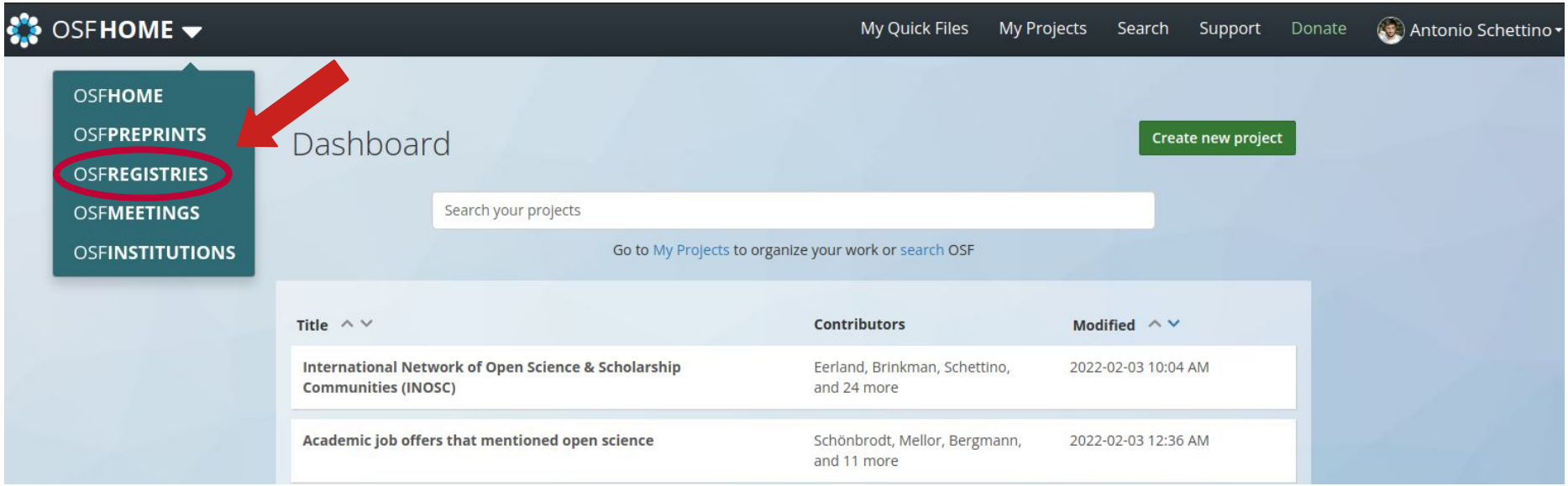
Log In

https://orcid.org/



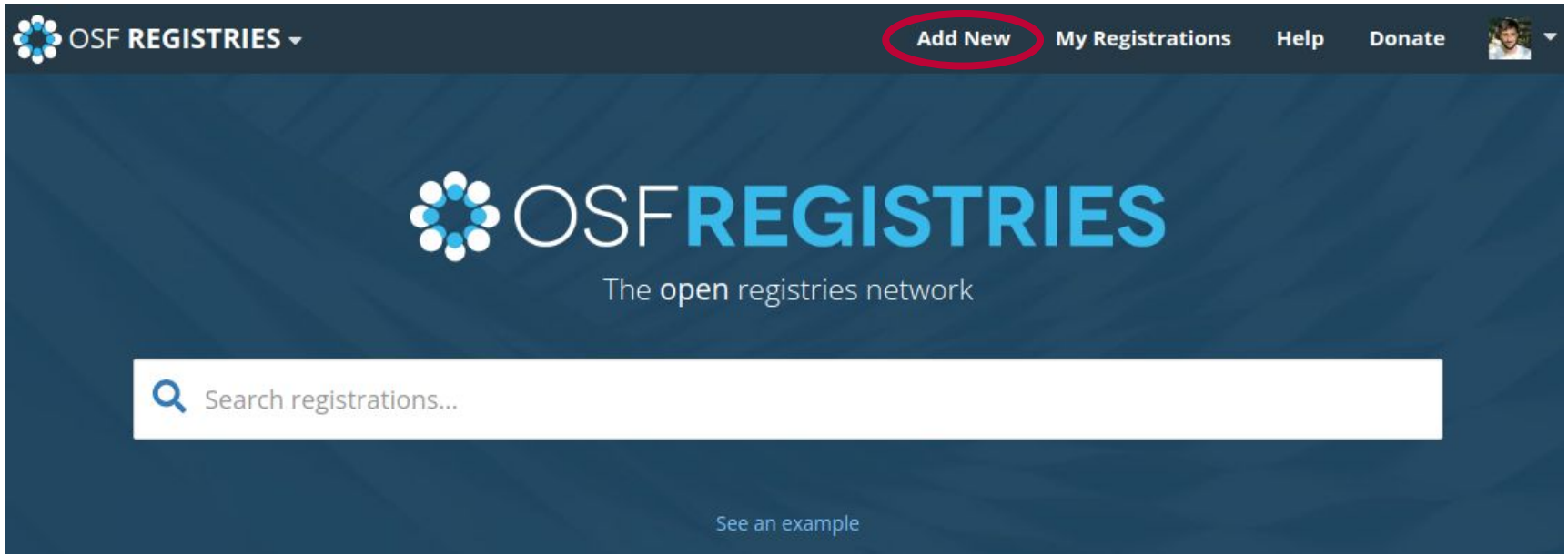
Antonio Schettino

OSF Dashboard



Antonio Schettino

OSF Registries



Antonio Schettino

Create First Draft

OSF REGISTRIES ▾ Add New My Registrations Help Donate

Add New Registration

You are submitting to OSF Registries. Click here to learn more about other hosted registries.

STEP 1

Do you have content for registration in an existing OSF project?

YES NO

STEP 2

Which type of registration would you like to create? *

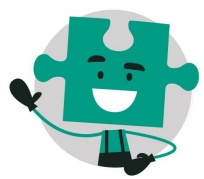
OSF Preregistration

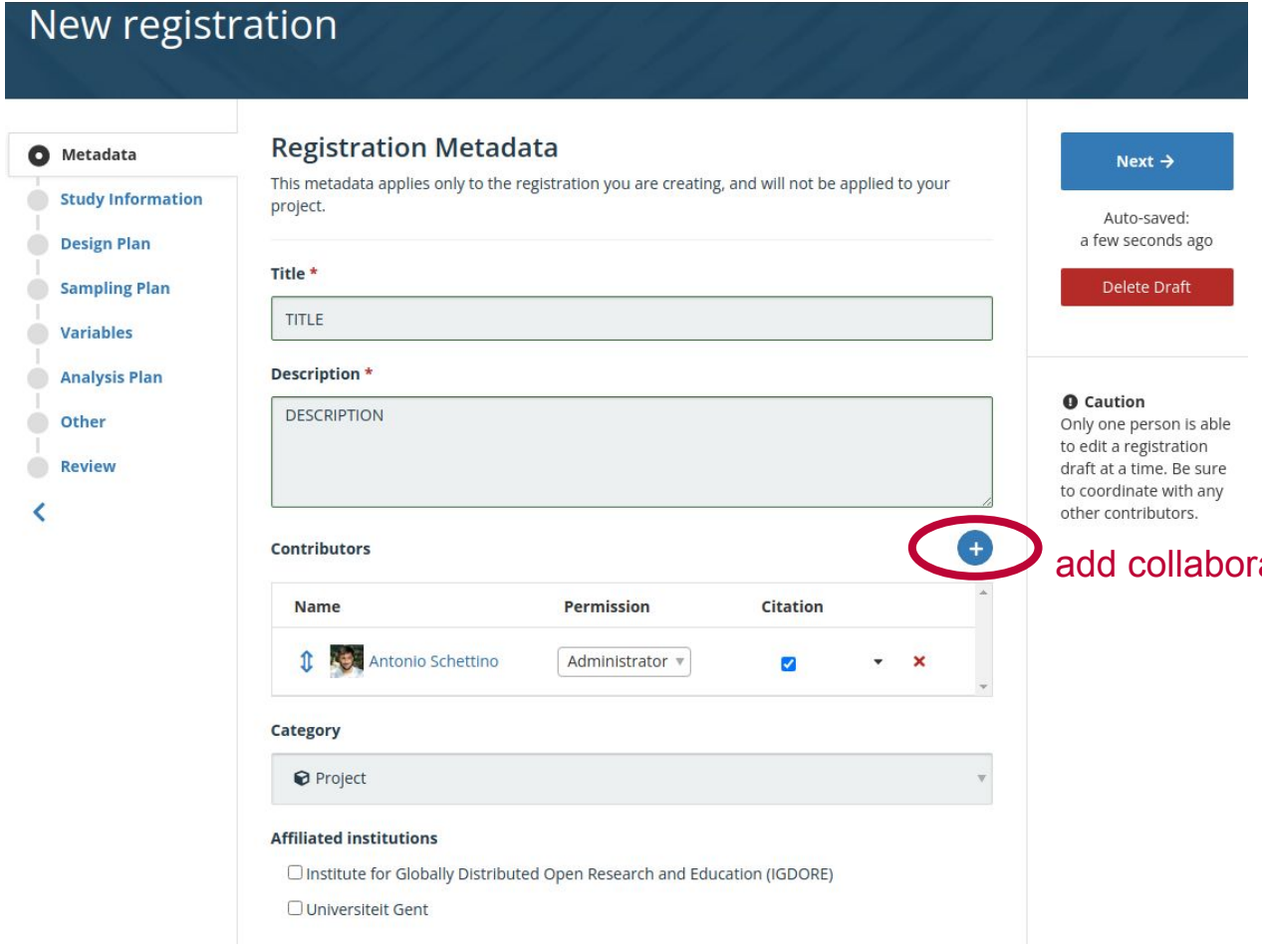
Create draft

- (•) OSF Preregistration
- ( ) Open-Ended Registration
- ( ) Qualitative Preregistration
- ( ) Registered Report Protocol Preregistration
- ( ) OSF-Standard Pre-Data Collection Registration
- ( ) Preregistration Template from AsPredicted.org
- ( ) Replication Recipe (Brandt et al., 2013): Post-Completion
- ( ) Replication Recipe (Brandt et al., 2013): Pre-Registration
- ( ) Pre-Registration in Social Psychology (van 't Veer & Giner-Sorolla, 2016): Pre-Registration

Antonio Schettino

Metadata

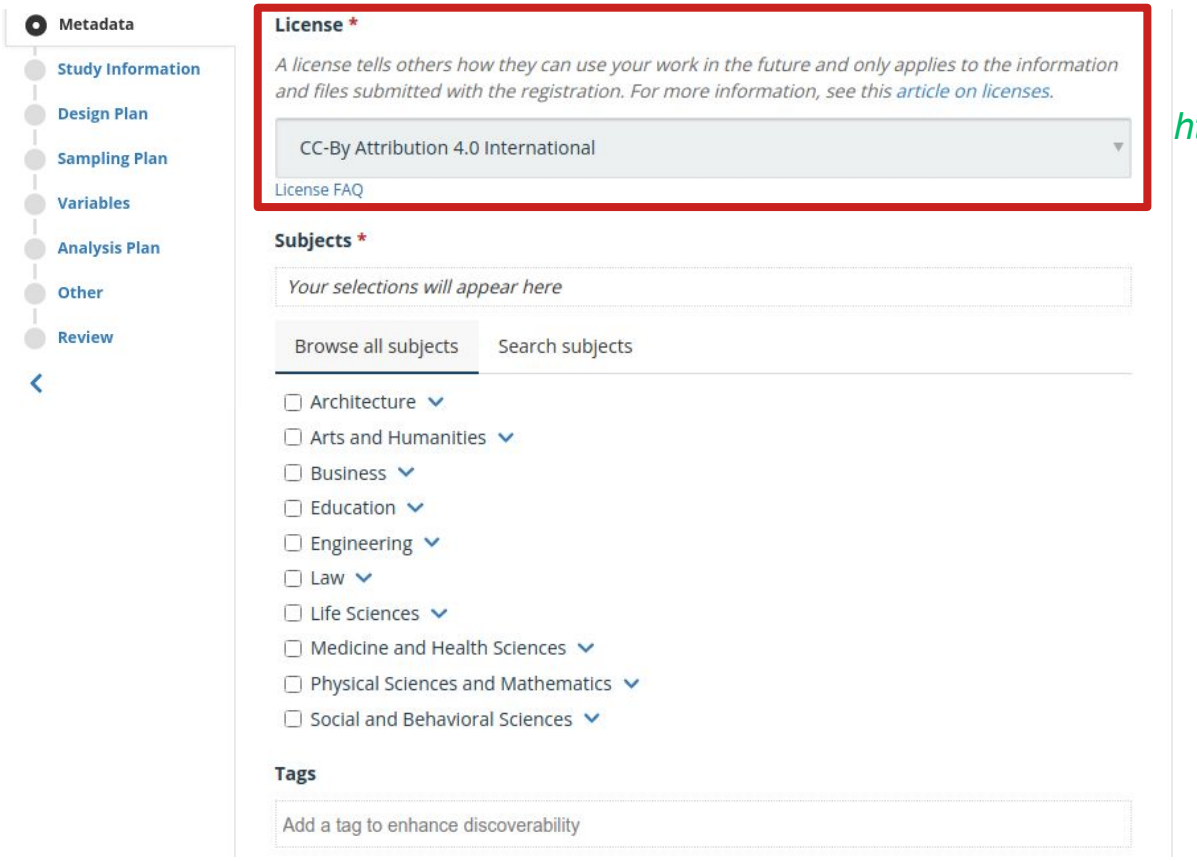




add collaborators

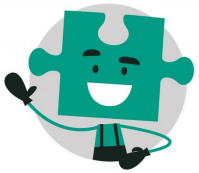
Antonio Schettino

Metadata (cont.)



see also
https://choosealicense.com

Antonio Schettino



Study Information

- Metadata
- Study Information
- Design Plan
- Sampling Plan
- Variables
- Analysis Plan
- Other
- Review

Study Information

Hypotheses *

List specific, concise, and testable hypotheses. Please state if the hypotheses are directional or non-directional. If directional, state the direction. A predicted effect is also appropriate here. If a specific interaction or moderation is important to your research, you can list that as a separate hypothesis.

Hide example

If taste affects preference, then mean preference indices will be higher with higher concentrations of sugar.

This field can't be blank.

Next →

← Metadata

Auto-saved:
2 minutes ago

Delete Draft

Caution
Only one person is able to edit a registration draft at a time. Be sure to coordinate with any other contributors.

Antonio Schettino

Design Plan

- Metadata
- Study Information
- Design Plan
- Sampling Plan
- Variables
- Analysis Plan
- Other
- Review

Design Plan

Study type *

Please check one of the following statements

- ○ Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.
- ○ Observational Study - Data is collected from study subjects that are not randomly assigned to a treatment. This includes surveys, "natural experiments," and regression discontinuity designs.
- ○ Meta-Analysis - A systematic review of published studies.
- ○ Other

Blinding *

Blinding describes who is aware of the experimental manipulations within a study. Mark all that apply.

- ☐ No blinding is involved in this study.
- ☐ For studies that involve human subjects, they will not know the treatment group to which they have been assigned.
- ☐ Personnel who interact directly with the study subjects (either human or non-human subjects) will not be aware of the assigned treatments. (Commonly known as "double blind")
- ☐ Personnel who analyze the data collected from the study are not aware of the treatment applied to any given group.

Is there any additional blinding in this study?

Blinding (Other)

Next →

← Back

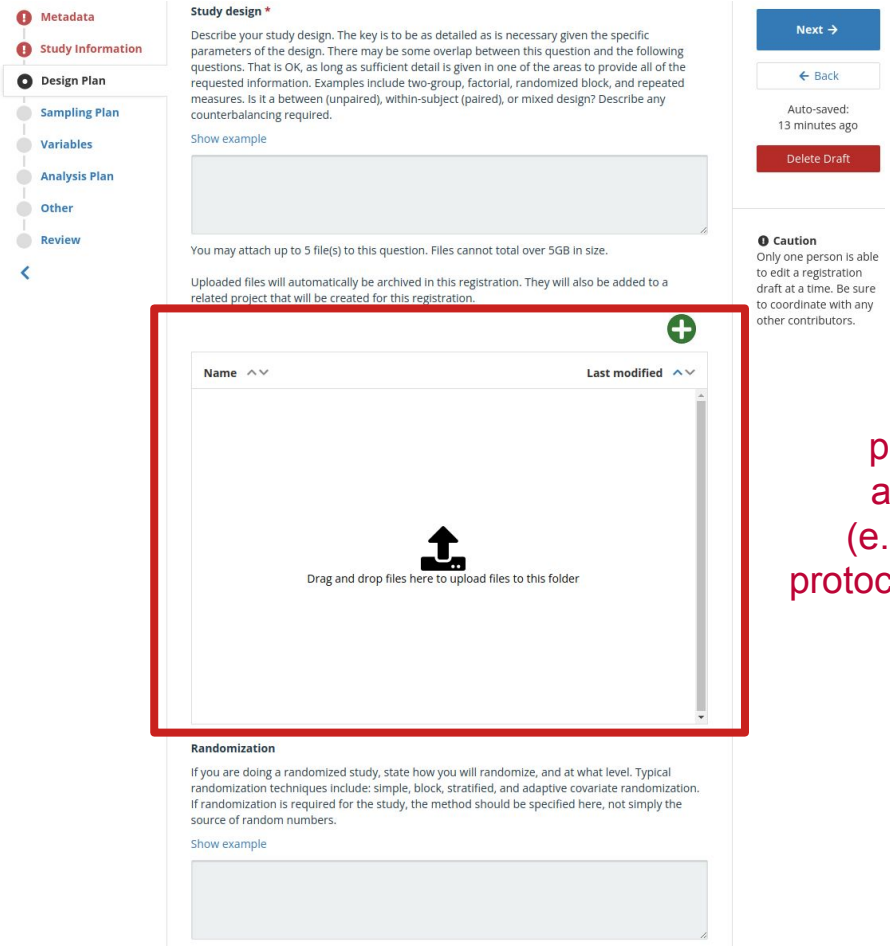
Auto-saved:
12 minutes ago

Delete Draft

Caution
Only one person is able to edit a registration draft at a time. Be sure to coordinate with any other contributors.

Antonio Schettino

Design Plan (cont.)



possibility to upload additional materials (e.g., scripts, interview protocols, questionnaires, …)

Antonio Schettino

Sampling Plan

- Metadata
- Study Information
- Design Plan
- Sampling Plan
- Variables
- Analysis Plan
- Other
- Review

Sampling Plan

Existing Data *

Preregistration is designed to make clear the distinction between confirmatory tests, specified prior to seeing the data, and exploratory analyses conducted after observing the data. Therefore, creating a research plan in which existing data will be used presents unique challenges. Please select the description that best describes your situation. See https://cos.io/prereg for more information.

- ○ Registration prior to creation of data
- ○ Registration prior to any human observation of the data
- ○ Registration prior to accessing the data
- ○ Registration prior to analysis of the data
- ○ Registration following analysis of the data

better use *preregistration template for secondary analysis*

not recommended

Explanation of existing data

If you indicate that you will be using some data that already exist in this study, please describe the steps you have taken to assure that you are unaware of any patterns or summary statistics in the data. This may include an explanation of how access to the data has been limited, who has observed the data, or how you have avoided observing any analysis of the specific data you will use in your study.

Next →

← Back

Auto-saved:
a few seconds ago

Delete Draft

Caution
Only one person is able to edit a registration draft at a time. Be sure to coordinate with any other contributors.

Antonio Schettino

Sampling Plan (cont.)

- Metadata
- Study Information
- Design Plan
- Sampling Plan
- Variables
- Analysis Plan
- Other
- Review

Data collection procedures *

Please describe the process by which you will collect your data and your inclusion and exclusion criteria. If you are using human subjects, this should include the population from which you obtain subjects, recruitment efforts, payment for participation, how subjects will be selected for eligibility from the initial pool, and your study timeline. For studies that don't include human subjects, include information about how you will collect samples, duration of data gathering efforts, source or location of samples, or batch numbers you will use.

Show example

You may attach up to 5 file(s) to this question. Files cannot total over 5GB in size.

Uploaded files will automatically be archived in this registration. They will also be added to a related project that will be created for this registration.

Name | Last modified

Drag and drop files here to upload files to this folder

Next →

← Back

Auto-saved:
a few seconds ago

Delete Draft

Caution
Only one person is able to edit a registration draft at a time. Be sure to coordinate with any other contributors.

possibility to upload additional materials

Antonio Schettino

Sampling Plan (cont.)

Sample size *

Describe the sample size of your study. How many units will be analyzed in the study? This could be the number of people, birds, classrooms, plots, or countries included. If the units are not individuals, then describe the size requirements for each unit. If you are using a clustered or multilevel design, describe how many units are you collecting at each level of the analysis. This might be the number of samples or a range, minimum, or maximum.

Show example

Sample size rationale

This could include a power analysis or an arbitrary constraint such as time, money, or personnel.

Show example

Stopping rule

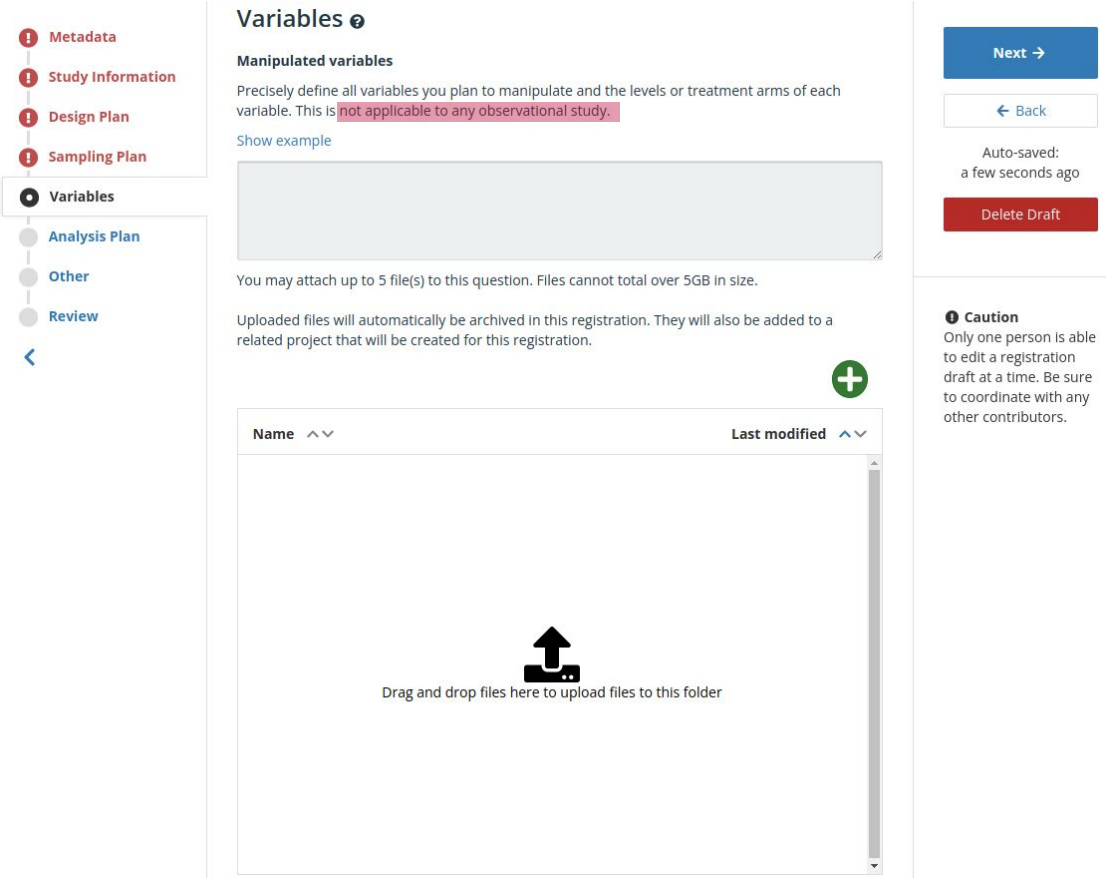
If your data collection procedures do not give you full control over your exact sample size, specify how you will decide when to terminate your data collection. If you are using sequential analysis, include your pre-specified thresholds.

Show example

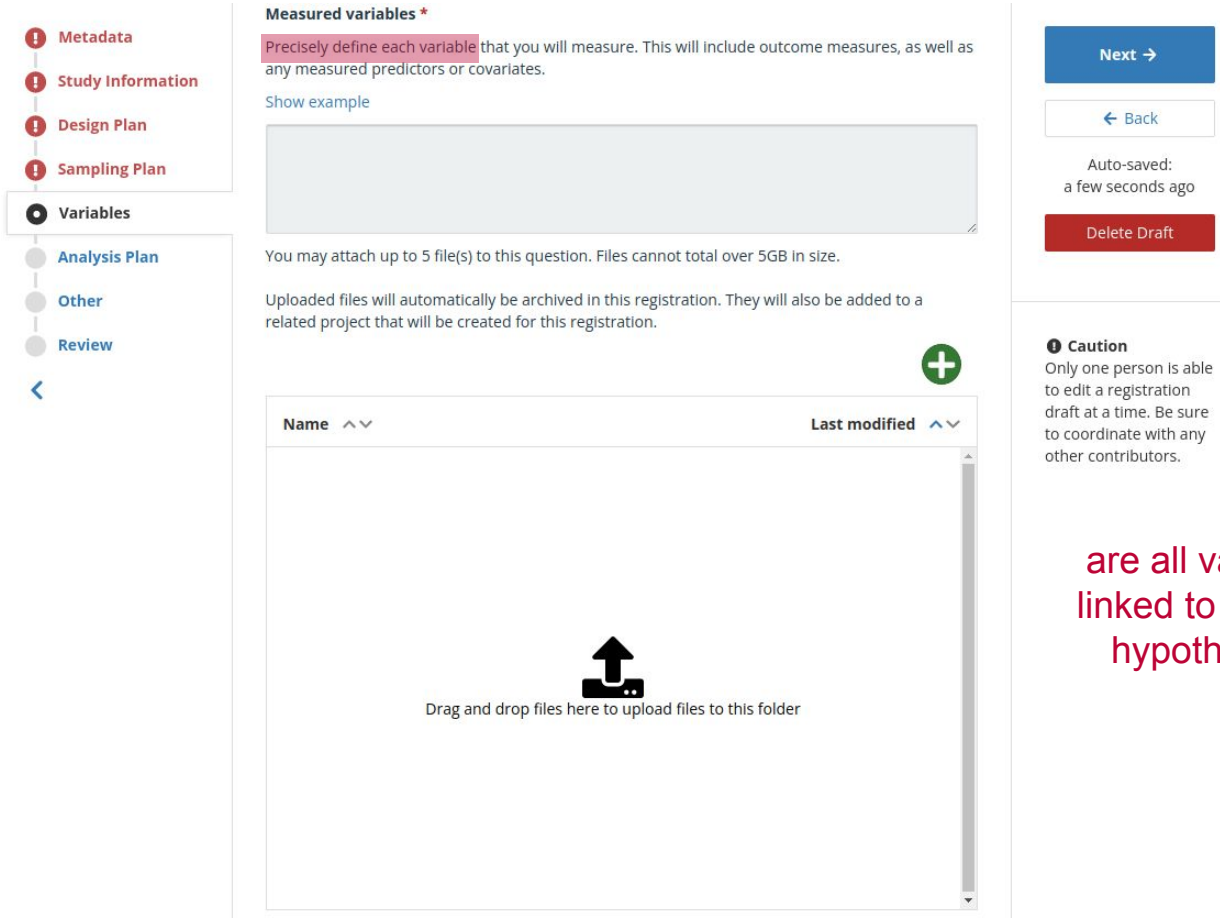
while *not all justifications have equal value*,
the most important thing
is **transparency**

Antonio Schettino

Variables



Variables (cont.)



are all variables linked to testable hypotheses?

Antonio Schettino

Variables (cont.)

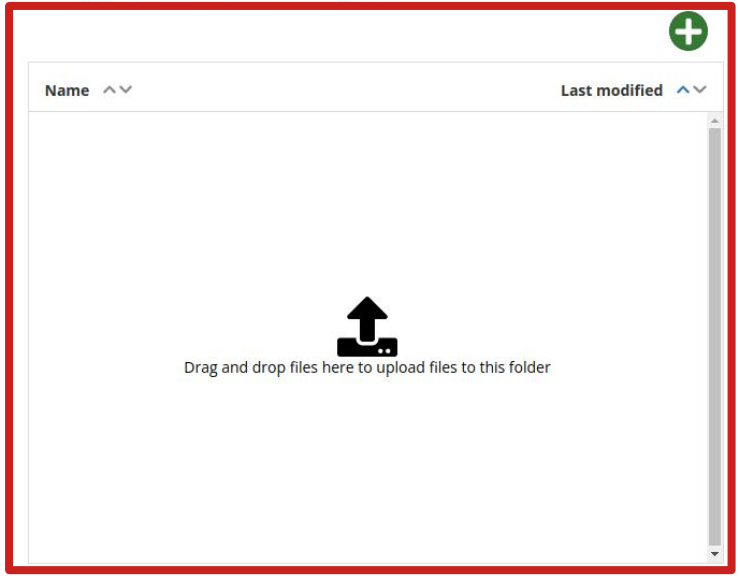
Indices

If applicable, please define how measures will be combined into an index (or even a mean) and what measures will be used. Include either a formula or a precise description of the method. If you are using a more complicated statistical method to combine measures (e.g. a factor analysis), please note that here but describe the exact method in the analysis plan section.

Show example

You may attach up to 5 file(s) to this question. Files cannot total over 5GB in size.

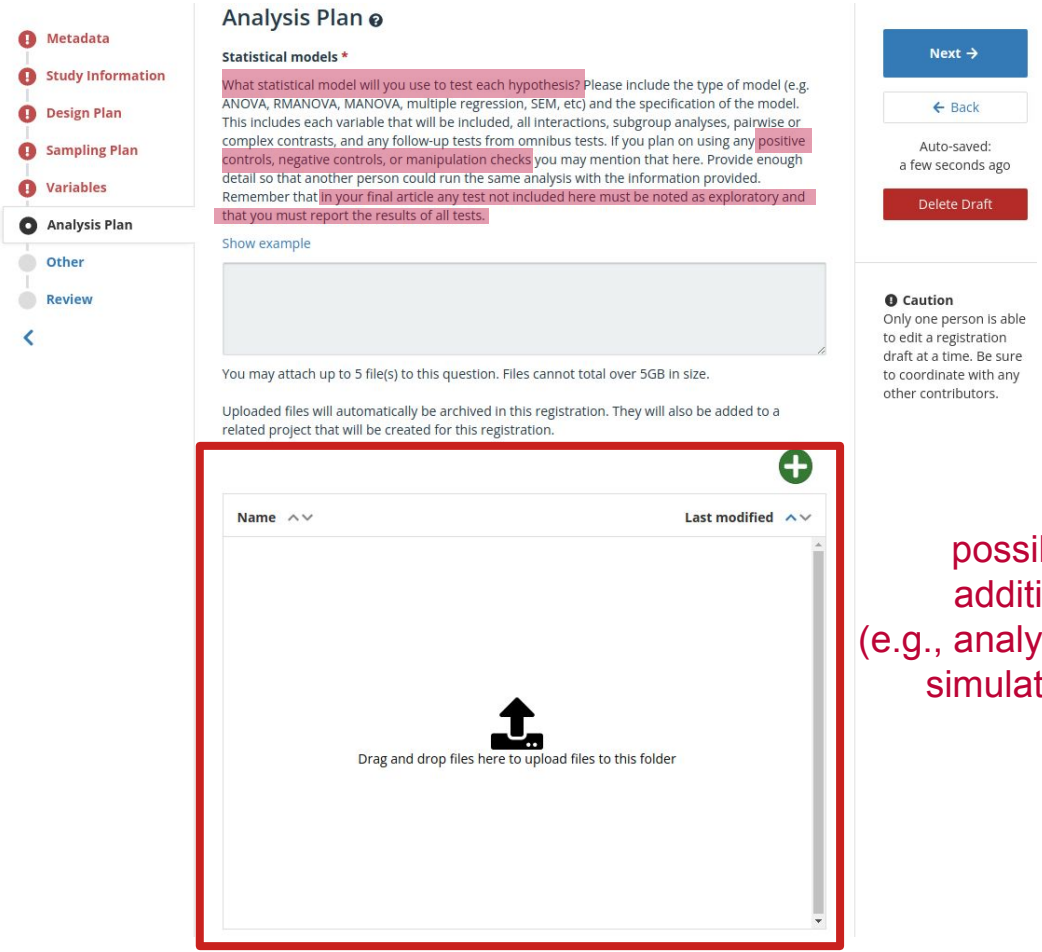
Uploaded files will automatically be archived in this registration. They will also be added to a related project that will be created for this registration.



possibility to upload additional materials (e.g., code for transformations)

Antonio Schettino

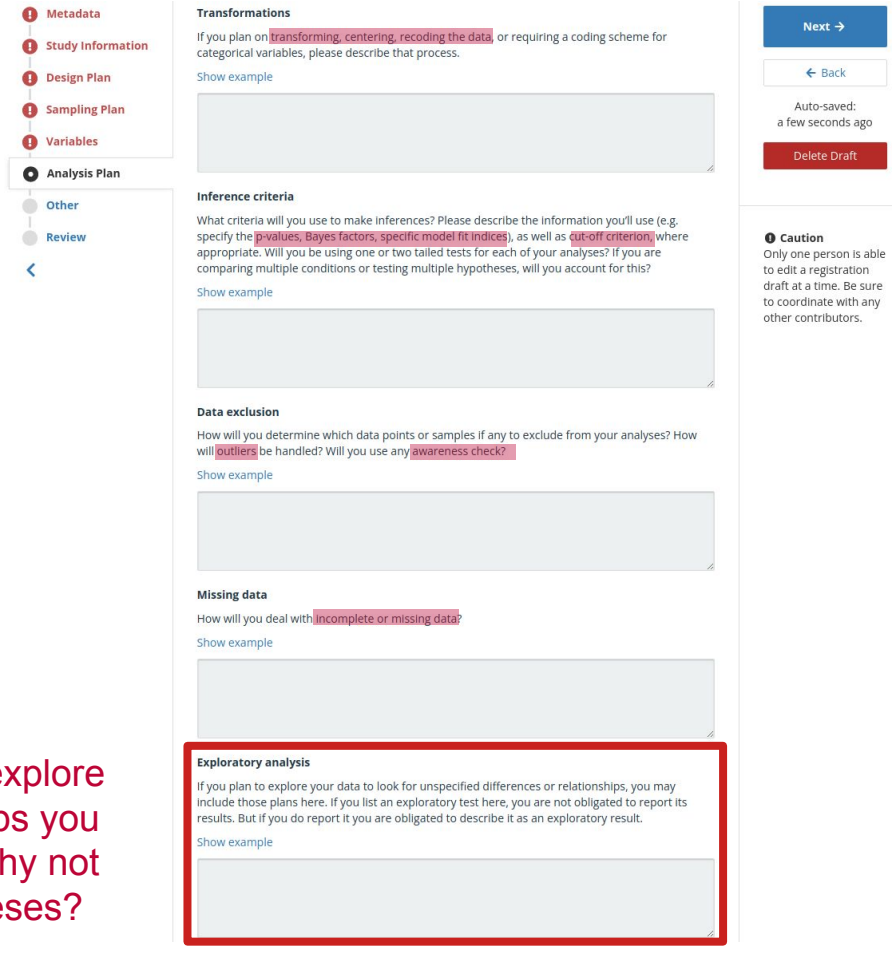
Analysis Plan



possibility to upload additional materials (e.g., analysis scripts tested on simulated or pilot data)

Antonio Schettino

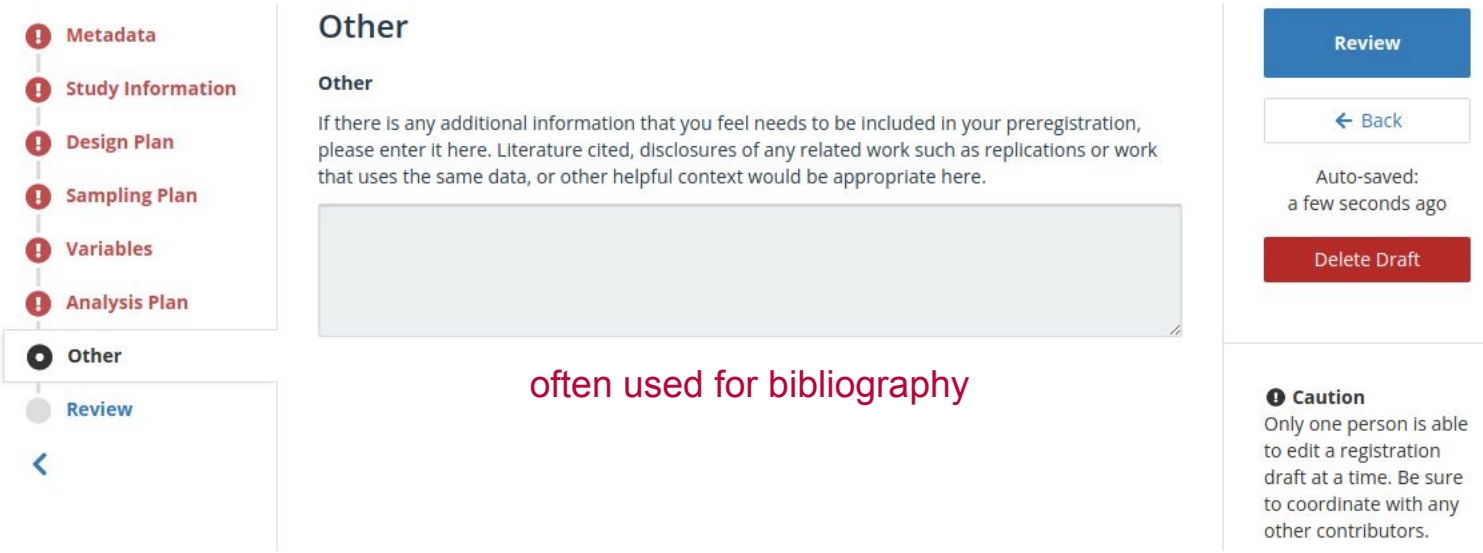
Analysis Plan (cont.)



if you already *plan* to explore some variables, perhaps you assume effects… so why not have testable hypotheses?

Antonio Schettino

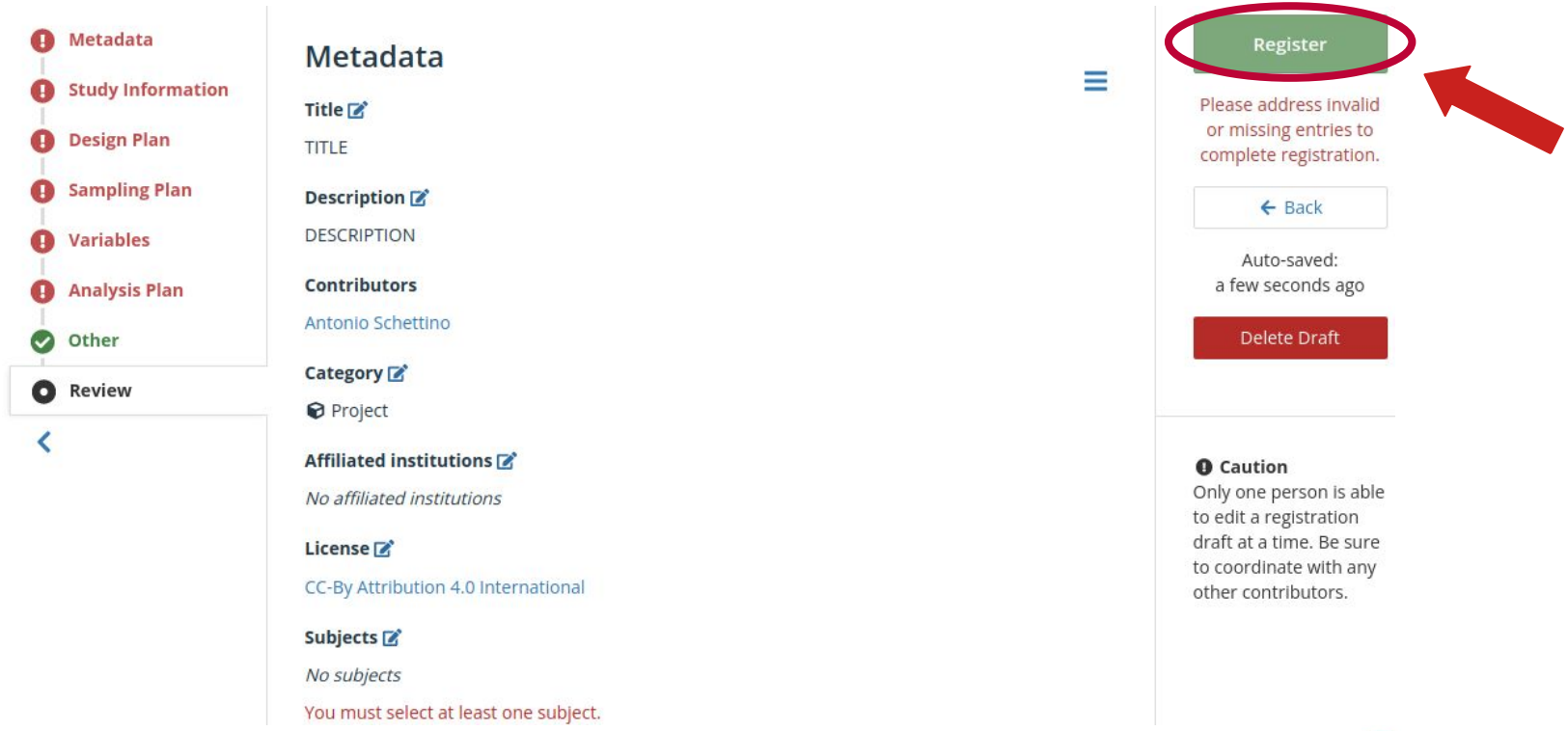
Other



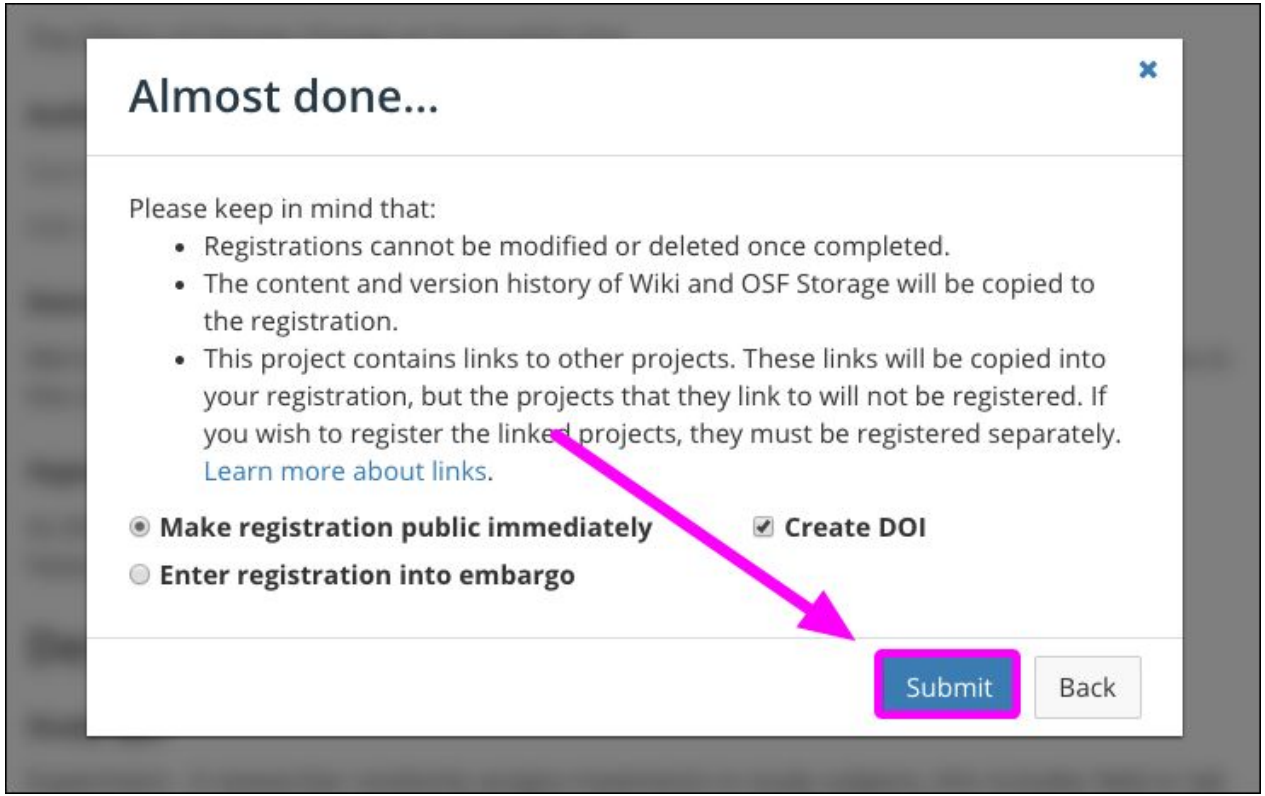
often used for bibliography

Antonio Schettino

Review



Register



Contributors have **48 hours** to approve/cancel the submission

Preregistration is published (either immediately or embargoed) if:

- all admin contributors have approved the submission

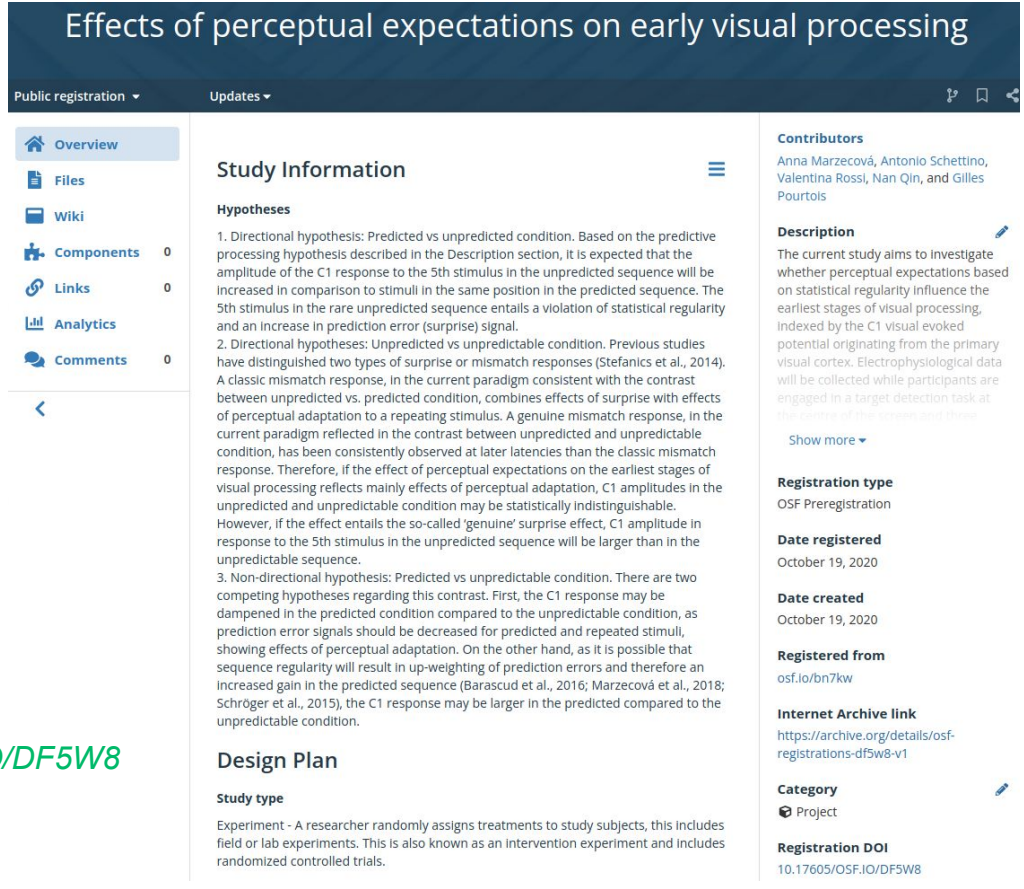
OR

- 48 hours have passed

https://help.osf.io/hc/en-us/articles/1500005374982-Submit-Your-Draft-Registration

Antonio Schettino

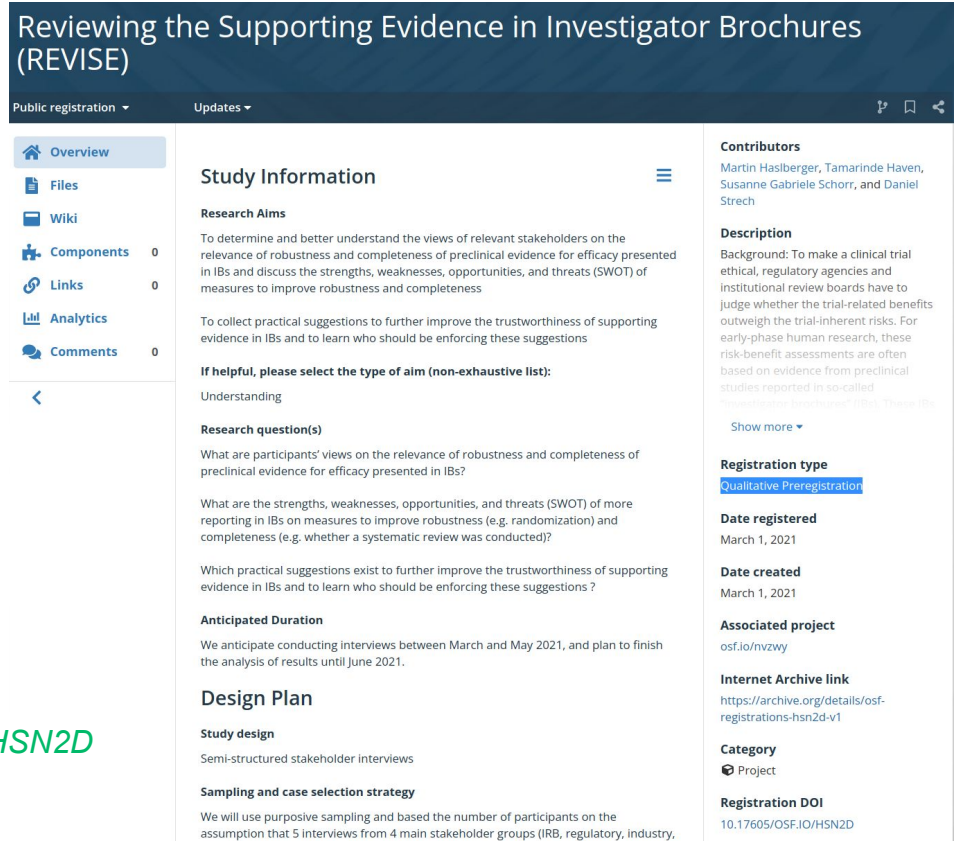
Example 1:
OSF Prereg Template



https://doi.org/10.17605/OSF.IO/DF5W8

Antonio Schettino

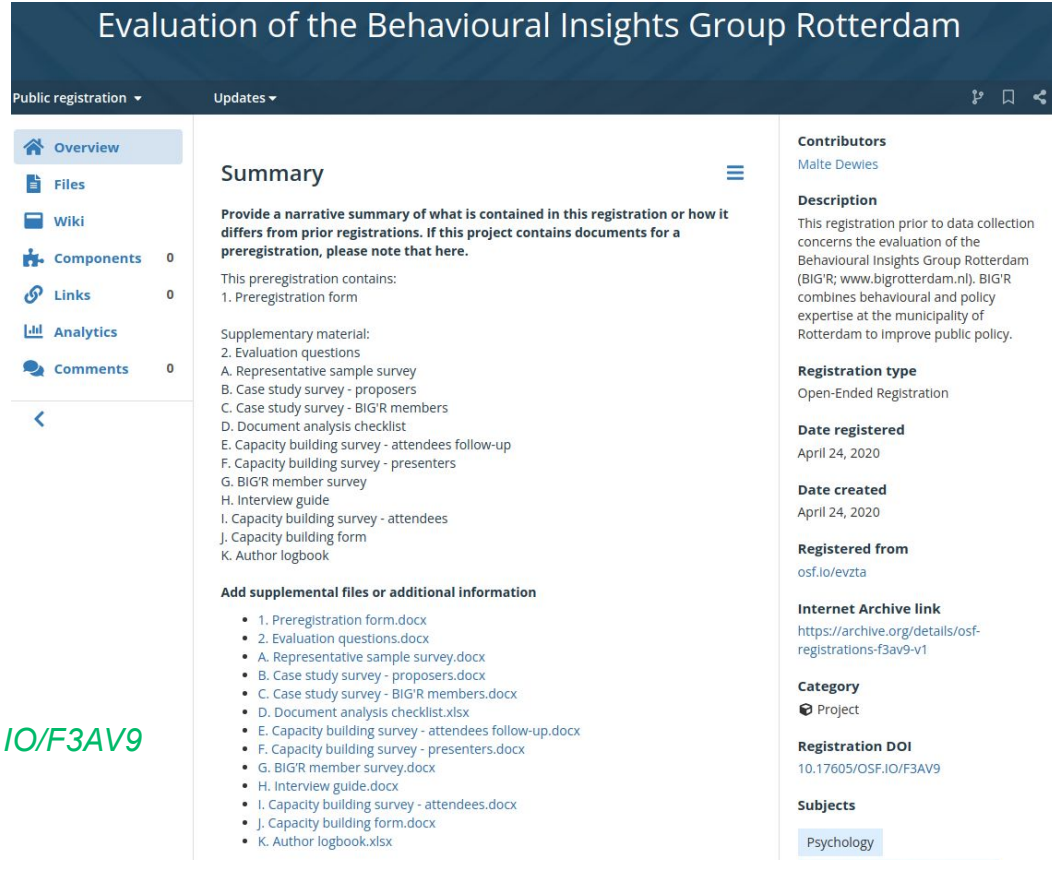
Example 2:
Qualitative Prereg Template



https://doi.org/10.17605/OSF.IO/HSN2D

Antonio Schettino

Example 3:
Open-Ended Prereg Template



https://doi.org/10.17605/OSF.IO/F3AV9

Antonio Schettino

Additional Information

- *Update* a preregistration
- Create *view-only (anonymized) link* to a preregistration
- *End an embargo* earlier than planned
- *Withdraw* a preregistration

Useful Resources

Literature

- *Preregistration: A pragmatic tool to increase transparency, reduce bias, and calibrate confidence in scientific research (Hardwicke & Wagenmakers, 2021)*
- *What should a preregistration contain? (McPhetres, 2020)*

Templates

- *Preregistration Templates*
- *Checklist* for comprehensive report of results
- Transparent Changes *Template Document*

OSF

- *OSF Guides* on registration
- *YouTube* tutorials

Antonio Schettino

HAPPY PREREGISTRATION!

Eliminating researcher degrees of freedom through Preregistration As Code (PAC)

Caspar J. Van Lissa[1], Aaron Peikert[2], Andreas M. Brandmaier[2]

[1]: Utrecht University, dept. Methodology & Statistics
[2]: Center for Lifespan Psychology, Max Planck Institute for Human Development

Defining preregistration

Specifying your research plan in advance of your study and submitting it to a registry.

Goal: Separating hypothesis-generating (exploratory) from hypothesis-testing (confirmatory) research.

Subgoals:

- Planning tool
- Improve quality and transparency
- Clearly report your study
- Set boundaries for agreed-upon work (PhD student, statistical collaborator)

Working reproducibly

WORCS: A Workflow for Open Reproducible Code in Science
www.developmentaldatascience.org/worcs

1. Dynamic document generation
   - Rmarkdown paper includes prose and analysis code
   - Paper can be reproduced / results updated with one click
2. Version control
   - Git tracks all changes since start of project
   - Project can be public on GitHub
   - Tag “release”: Time capsule at specific stage of project, e.g. “pre-registration”
3. Dependency management
   - Record or container with all software needed to reproduce

Preregistration in WORCS

1. Complete a prereg template in Rmarkdown format `prereg.Rmd`
2. Commit and push to GitHub
3. Tag the release as `preregistration`
4. Optional: Render Rmarkdown to PDF and upload to OSF.io / aspredicted.org
5. Collect data
6. Write `manuscript.Rmd` with planned analyses

Shortcomings of preregistrations

- Extra work
- Residual ambiguity
  - Unintentional
  - Strategic
- Straight-jacket: Can “force” researcher to stick with bad plans
- Researchers often not trained in preregistration
  - In neither writing nor reviewing thereof
- Not all preregistration templates are relevant for all research (e.g., secondary analysis)
- Difficult to compare with final manuscript, because they are in different formats

Solution: PAC

1. Preregister Rmarkdown with draft of manuscript
2. Include code for planned analyses
3. Use fake* data to obtain mock results
4. Draft report based on mock results

After collecting real data

1. Re-compile Rmarkdown and see final results
2. Write Results and Discussion

* = Simulated, or synthetic, or from prior study, or shuffle some real data!

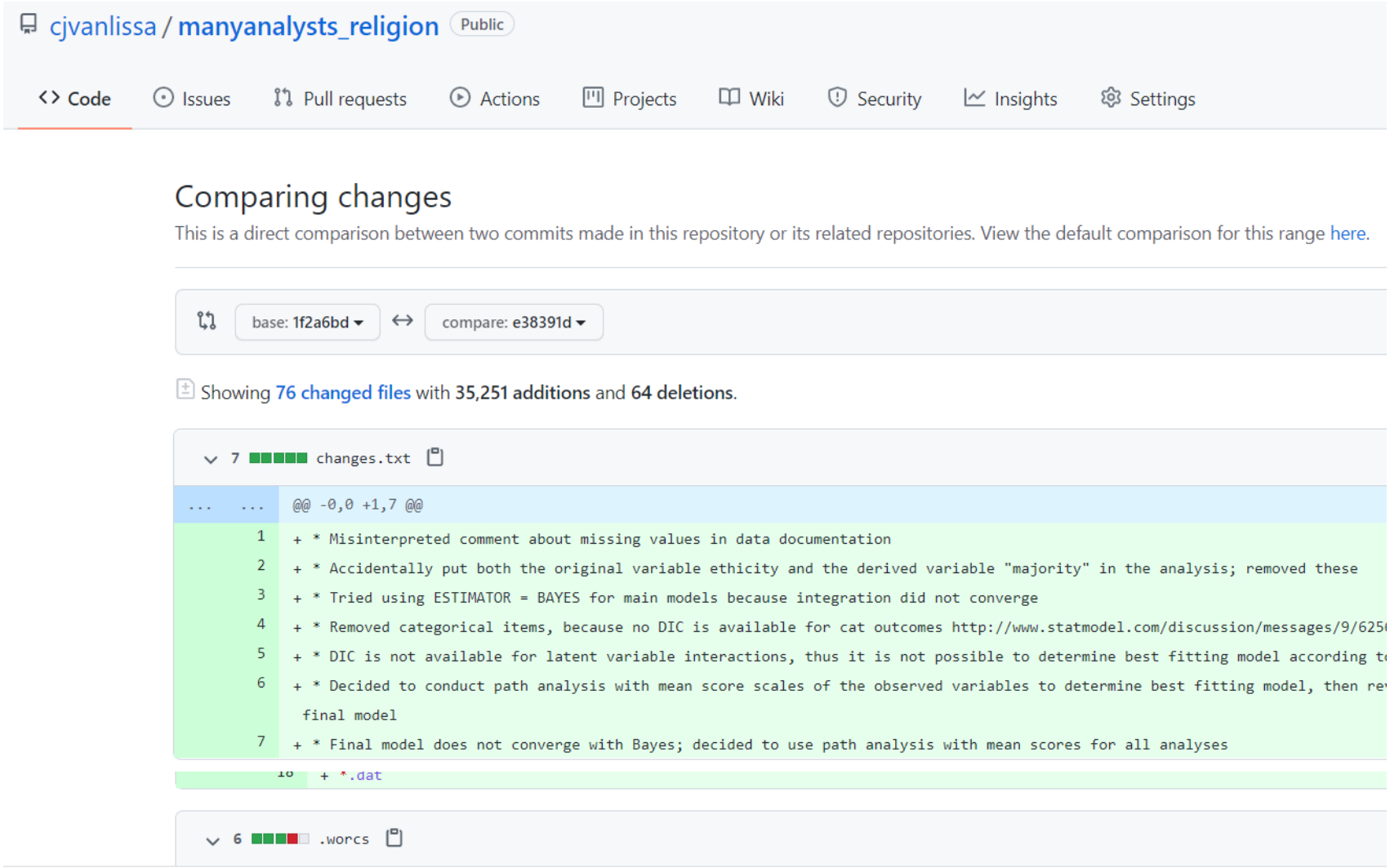
Advantages

- Less work
  - You're not writing a separate document, but an early version of the final manuscript
  - You need to write code to analyze the data anyway
- unambiguous
  - Unintentional ambiguity ↓: You become aware of ambiguity in your planned analysis when you start actually running it
  - Strategic ambiguity ↓: there's a straightforward game plan
- No Straight-jacket: Can't "force" researcher to stick with bad plans
  - You can deviate from planned analyses; create a Git commit with the changes

Advantages 2

- Researchers are trained in writing papers, not preregistrations! Sticking with that format is easier to write and review
- Writing a preregistration as draft manuscript ensures all sections are directly relevant
- You can literally compare the two versions (using Git diff) to see how the planned analyses were executed

Example


cjvanlissa / manyanalysts_religion
Public
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
Comparing changes
This is a direct comparison between two commits made in this repository or its related repositories. View the default comparison for this range here.
base: 1f2a6bd
compare: e38391d
Showing 76 changed files with 35,251 additions and 64 deletions.
7 changes.txt
@@ -0,0 +1,7 @@
+ * Misinterpreted comment about missing values in data documentation
+ * Accidentally put both the original variable ethicity and the derived variable "majority" in the analysis; removed these
+ * Tried using ESTIMATOR = BAYES for main models because integration did not converge
+ * Removed categorical items, because no DIC is available for cat outcomes http://www.statmodel.com/discussion/messages/9/625
+ * DIC is not available for latent variable interactions, thus it is not possible to determine best fitting model according t
+ * Decided to conduct path analysis with mean score scales of the observed variables to determine best fitting model, then re
final model
+ * Final model does not converge with Bayes; decided to use path analysis with mean scores for all analyses
+ *.dat
6 .worcs

Limitations

- Can’t cover all contingencies; deviations will be necessary
  - Version control makes clear what has changed
- Simulating data is difficult/requires much effort
  - But super useful, e.g. power analysis
  - Statistical co-author
  - Collect data first, shuffle dependent variable
- Requires reproducible/reusable workflow
  - But this ensures correctness/reliability/reusability of your work
- Preregistration forms may be more detailed / complement information in manuscript
  - What can we learn from prereg forms to improve our manuscripts? # Problem 2 — Writing a Preregistration

Conclusion

Preregistration As Code may have advantages

- Same format as paper
- Unambiguous
- Easy to compare prereg VS final version
- Power analysis etc

But… they require training in reproducibility and statistical programming that is not yet commonplace

So: A vision of the future of preregistration?