Center for Effective Global Action
Guidelines for Transparent and Reproducible Research

Contents

Summary

CEGA recognizes that for research to be deemed credible, it must be conducted in a transparent and reproducible manner. Espousing the principles of research transparency and reproducibility dictates that researchers clearly and precisely document, report, and share the data, materials, methods, and analytical decisions used during the course of a project in ways that facilitate replication, collaboration, and reuse.

Research transparency and reproducibility advances CEGA's mission in three important ways by:

1. Asserting the rigor of research used for policy and affirming its evidentiary value through replication;
2. Enabling cooperation, accumulation of knowledge, and free exchange of ideas among the research community; and
3. Facilitating the inclusion of scholars from low- and middle-income countries (LMICs) by lowering barriers to access and participation.

These guidelines articulate CEGA's commitment to high standards of transparency and reproducibility for all empirical research projects supported in whole or in part by CEGA. In cases where research projects supported by CEGA are subject to additional transparency and/or reproducibility policies – for example, from donor organizations – the higher standards necessarily apply.

Researchers supported by CEGA are **required** to:

1. Register all prospective research studies with a recognized registry before data collection for the project (this does not include developing pre-analysis plans). Preliminary fieldwork and pilot projects are exempt from this requirement.
2. Disclose in all proposal submissions to CEGA whether, and if so, how they intend to share all data, code and study materials for the purpose of replication and reuse.

We **recommend** that researchers also implement the following measures:

3. Develop pre-analysis plans, documenting in detail all procedures for data collection, processing, and analysis;
4. Ensure that research outputs are fully computationally reproducible, in that data, materials, code, and detailed accounts of study procedures would allow an informed researcher to reproduce all of the results precisely and with minimal effort
5. Provide unrestricted access to, and reuse of, all publications and research outputs resulting from projects, for example through open access publication.

Each of these measures is elaborated below, with details on best practices for implementation.

CEGA will provide public links to all replication materials (including pre-registrations, pre-analysis plans, data, study materials, and code) **in communications related to supported research projects**, including project descriptions on the CEGA website and technical reports. We will also disclose cases where there are legal or ethical restrictions on publishing any of the data or materials.

This is a living document to be continuously updated based on feedback from affiliated CEGA researchers and lessons learned in the process of implementation. We welcome feedback to Carson Cristiano, Executive Director at christiano@berkeley.edu.

I. Research design and analysis

A. Registration (required)

CEGA **requires** that Principal Investigators (PIs) register with a trusted registry all prospective empirical research projects for which CEGA provides funding. Projects need to be registered *before* data for that particular project are collected, with the exception of preliminary fieldwork and pilot projects. Researchers must provide to the responsible CEGA Project Manager a view-only link[1] to the registration before any funds are issued.

Researchers may embargo the registration in full (or redact sensitive details) to protect the integrity and confidentiality of ongoing studies. All registrations must be *made public* within 18 months of the end of data collection, or when a paper presenting the main results of the study is published, whichever occurs first.

How to:

CEGA recommends that researchers submit their registrations to the American Economic Association's registry for randomized controlled trials (Social Science Registry), though researchers may choose another registry, such as the Open Science Framework (OSF) registry, the Evidence in Governance and Politics (EGAP) registry, the Registry for International Development Impact Evaluations (RIDIE), AsPredicted.org, or ClinicalTrials.gov. See section IV. A. for more information.

Some registries require certain details of the design to be submitted by default, however registering a study still requires a low level of effort. We recommend that registrations, at a minimum, contain the following details: hypotheses, sample size and power calculations, basic study design (identification strategy), main outcome variables, control variables, intervention description (including start date and end date), and statistical model. Note that *registrations can be updated if any details or objectives circumstances of the study change.*

B. Pre-analysis plans

CEGA recommends that researchers also include a pre-analysis plan (PAP) in their registration, providing extensive methodological descriptions of the analysis *before any data are analyzed*[2]. Depositing a PAP to a trusted hypothesis registry generates a verifiable time-stamped confirmation that analyses were specified before being conducted.

How to:

Though there are no conventions, PAPs in economics typically contain the following details: *i)* study design (identification strategy); *ii)* study sample (including sample size, sampling frame, assignment to treatment, and corresponding statistical power calculations); *iii)* outcome measures (for RCTs, also a description of the intervention and randomization strategy); *iv*) mean effects family groupings; *v)* multiple hypothesis testing adjustments; *vi)* subgroup analyses; *vii)* direction of effect for one-tailed tests; *viii)* statistical specification and method; *ix)* structural model; and *x)* time stamp for verification.[3] See section IV. B. below for more information.

Exploratory analyses and serendipitous findings may still be reported, but should be clearly distinguished from pre-registered, confirmatory analyses. This helps preserve the integrity of both without favoring one over the other.

---

[1] See instructions on how to create a view-only link on the OSF registry while maintaining your registration hidden to the public.
[2] The timing of pre-analysis plans is still an issue of debate. The earlier a PAPs is developed (before baseline), the less room there will be for bias (an unintentional p-hacking). However, developing a PAP after the baseline (but before endline) allows researchers to be more specific about the planned analyses.
[3] Based on Christensen and Miguel (2018).

II. Data, code and research materials

A. Sharing data, code and research materials

CEGA **requires** that researchers **disclose** in all proposal submissions whether they intend to share all data, code and study materials for the purpose of replication and reuse. If researchers agree to share replication materials, they must specify where such materials will be available. Researchers must also disclose if legal or ethical restrictions apply to all or part of the data.

How to:

We recommend that all sharing is done through a trusted data repository[4] and no later than 18 months after the end of data collection or when the results of the study are published, if sooner. Researchers should provide to the responsible Project Manager a digital object identifier (DOI) or link to the repository containing data and study materials. See section IV. C. for an overview of trusted data repositories.

To facilitate replication and reuse of research, researchers should also consider the following:

i. Raw data[5] should be shared, such that the final results of a study can be reproduced in full.
ii. All data, program code, scripts, codebooks, and other documentation should be provided, including a description of the procedures necessary to conduct an independent replication of the research.
iii. Data for all variables, treatment conditions, and observations described in the manuscript should be shared.
iv. Consider using cloud computing services like Binder (open source) or Code Ocean (proprietary) to facilitate computational reproducibility and mitigate against version and licensing dependencies in statistical analysis software.

B. Data de-identification and ethics

Principal Investigators are required to ensure that all data from human subjects **is** fully de-identified before being shared publicly. CEGA expects that in all projects supported by the Center, PIs ensure compliance relevant international and national data privacy and protection laws, and protocols approved by an Institutional Review Board or equivalent governing body. CEGA has no obligation to monitor the content of any data, code, and study materials shared for the purpose of replication and reuse.

How to:

See section IV. C. for select relevant resources.

C. Data with limited or restricted access

In cases where legal, ethical or other circumstances preclude public access to the data and/or materials, researchers should disclose this fact in all technical reports, publications and other research communications.

How to:

We recommend that the following details be reported:

[4] Trusted repositories provide unique and persistent digital object identifiers (DOI), and adhere to policies that make data discoverable, accessible, usable, and preserved for the long term.
[5] Raw data is any data that was collected by the researcher and has not been processed in any way, other than the removal of personal identifiers.

i. An explanation of the restrictions on the dataset or materials and how they preclude public access;
ii. A public description of the steps others should follow to request access to the data or materials, including specific contact information and the unique identifier[6] of the data set;
iii. Software and other documentation that will precisely reproduce all published results; and
iv. Access to all data and materials for which the constraints do not apply, or consider sharing a processed data set from which the final analyses can be reproduced.

D. Reproducibility

Researchers should ensure that all results communicated in all research outputs resulting from supported research projects are fully computationally reproducible, in that *data, materials, code, and detailed accounts of study procedures would allow an informed researcher to reproduce all of the results precisely and with minimal effort*.

How to:

We recommend that researchers adopt best practices for computational reproducibility, including the use of literate programming (coding and commenting code in such a way that humans can understand how a machine analyzes the data), modular programming (separating data cleaning, processing, recoding, and merging from analysis), preparing codebooks and readme files, and using version control software (e.g. Git, Mercurial, etc.).

In light of CEGA's commitment to capacity building for developing country scholars, we also recommend that researchers post replication materials that are reproducible through *open source* software, such as R or Python.

III. Scholarly communication

A. Open access

In an effort to minimize barriers to knowledge and to facilitate the inclusion of developing country scholars, CEGA recommends that researchers publish in open access publications, or minimize restrictions to research materials and publications to the furthest extent possible.

How to:

Researchers should choose journals that allow for open access publishing and opt for *gold open access* (publication is available immediately on the journal's website and is available to readers free of charge). Open access publishing fees may be built into award budgets, unless fees are already covered by a donor organization (e.g. projects funded by the Gates Foundation) or affiliates' institutions.

B. Preprints

Where publishing open access is not possible, CEGA encourages researchers to archive their research outputs using preprints (open access versions of scholarly works).

How to:

Preprint services are *free of charge*, and accept submissions in the form of working papers, full papers (both pre- and post-publication), technical reports, tutorials for statistical software packages, and conference proceedings.

[6] The unique identifiers are obtained by applying hash functions (for example, 'datasignature()' in Stata and 'digest()' in R) to the original raw data and obtaining a long alphanumeric string that cannot be reversed.

Note that preprints are the preferred to self-archiving on personal or organizational websites because they are searchable, long-lasting, and provide stable DOIs.

For a list of relevant preprint services, as well as publisher copyright and self-archiving (preprint) policies, see section IV. F.

C. Registered Reports (pre-results review)

CEGA encourages researchers to consider publishing in journals offering Registered Reports (pre-results review), where papers are reviewed and accepted based on their methods and theory before the results of the study are known. Pre-results review submissions are similar to PAPs, allowing researchers to use an existing PAP to prepare a pre-results review submission with minimal effort.

How to:

CEGA supported the *Journal of Development Economics* in a pilot implementation of this format and more than 160 journals across the social and life sciences currently accept registered reports. Find Author Guidelines and other helpful resources on the BITSS website.

IV. Resources[7]

A. Registration

The table below describes the most popular study registries in the social sciences:

Registry	Type of research	Embargo	Notes
AEA Registry	RCTs in non-health-related fields	Until completion of the study	Default for RCTs in economics
EGAP registry	RCTs and observational studies in governance and politics	Up to 18 months after registration	Popular in political science
Clinical Trials	RCTs in health-related fields	Undefined limit	Highly standardized
Registry for International Development Impact Evaluations (RIDIE)	Impact Evaluation, Development Economics	Unavailable	High level of detail. Accepts RCTs and quasi-experimental.
OSF Preregistration	Any	Up to 4 years after registration	Multiple formats: short, long, structured, and open ended

A post on the Data Colada blog features examples of good practices for registering various aspects of a research design.

B. Pre-analysis plans

Below is a list of pre-analysis plan templates and guidelines:

- Ganimian, Alejandro. "Pre-Analysis Plan Template." n.d. Link.
- *Journal of Development Economics.* "Stage 1 Proposal Template." Berkeley Initiative for Transparency in the Social Sciences (BITSS), 2018. Link.
- Food and Drug Administration. "Guidance for Industry, E9 Statistical Principles for Clinical Trials." US Department of Health and Human Services, 1998. Link.
- McKenzie, David. "A Pre-Analysis Plan Checklist." Impact Evaluations, 2012. Link.
- Burlig, Fiona. "Improving Transparency in Observational Social Science Research: A Pre-Analysis Plan Approach." *BITSS Preprints*, October 30, 2017. Link.

[7] This section is based on resources curated in the Inter-American Development Bank's "Best Practices for Transparent, Reproducible, and Ethical Research", developed by Fernando Hoces de la Guardia (BITSS) and Jennifer Sturdy (Millennium Challenge Corporation).

C. Data sharing

The following repositories issue DOIs and meet the highest standards for data access, preservation, and stability:

Repository	Fees and costs	Size limits
Dryad Digital Repository	120USD for the first 20 GB, and 50USD for each additional 10 GB	None stated
figshare	100 GB free per *Scientific Data* manuscript. Additional fees apply for larger datasets	1 TB per dataset
Harvard Dataverse	Contact repository for datasets over 1 TB	2.5 GB per file, 10 GB per dataset
Open Science Framework	Free of charge	5 GB per file, multiple files can be uploaded
Zenodo	Donations towards sustainability encouraged	50 GB per dataset
Mendeley Data	Contact repository for datasets over 10 GB	10 GB per dataset
ICPSR	Based on institutional membership	None stated

D. Data management and de-identification

IPA's "Best Practices for Reproducible Research" (2015) is a concise guidebook on best coding practices for reproducibility in Stata. Link.

The World Bank's DIME Wiki contains a range of useful resources for data management and de-identification, including a Checklist for Data Cleaning, Stata Coding Practices, and Publishing Data.

J-PAL's "Data Security Procedures for Researchers" is a primer in a range of data security topics relevant for impact evaluations. Link.

The ARDC FAIR Data self-assessment tool allows researchers to evaluate whether their data is findable, accessible, interoperable, and reusable.

"Practical Tips for Ethical Data Sharing" by Michelle N. Meyer (2018) is a practical tutorial with do's and don'ts for data sharing. Link.

The 2018 International Compilation of Human Research Standards by the U.S. Department of Health and Human Services lists over 1,000 laws, regulations, and guidelines on human subjects protections in 130 countries and from many international organizations.

Data Protection Laws of the World by DLA Piper Law Group and Data Protection around the World by Commission Nationale de l'Informatique et des Libertés (CNIL) allow users to compare laws and regulations between countries.

E. Reproducibility

Using a coding style guide, authors can make it easier for collaborators and replicators to review a script and find relevant pieces. Examples include:

- Google's R style guide
- General style guide for Python
- The Stata Journal style guide

Jake Bower's "Six Steps to a Better Relationship with Your Future Self" is a short, practical guide on best practices for reproducible workflows. Link.

F. Scholarly communication

SHERPA RoMEO is a database of publisher copyright and self-archiving (preprint) policies. Consult this database when deciding to post your paper on a preprint service.

ArXiv is a general purpose preprint service originally used in mathematics and physics, but has gained traction among researchers in economics.

EconStor is an open access repository specifically intended for economic literature, allowing authors to submit papers free of charge.

SSRN, particularly the Economics Research Network, allows authors to post and download papers for free.

The Center for Open Science maintains a hub of resources for registered reports, including a list of participating journals, FAQs, and templates.