Guidelines for Investments in Research

At the Laura and John Arnold Foundation (LJAF), we are committed to funding research that meets the most rigorous standards of quality and transparency. These Guidelines establish what is expected of LJAF's research grantees and consultants, subject to modification to address a researcher's particular needs or circumstances. We will work with grantees and researchers to tailor any disclosure requirements so that all confidentiality and privacy requirements are respected. Section I describes the process of preregistration. Section II discusses how and when computer code and datasets should be made public.

I. Preregistration

One of LJAF's key efforts to address publication bias and improve the reliability of empirical analysis is to require that research plans be preregistered. In brief, preregistration involves describing the research design in detail *before* the statistical analyses are performed and, if possible, before data are collected. We require preregistration via the Open Science Framework (OSF), hosted by the Center for Open Science.[1] The OSF is a web-based platform that scientists can use to streamline and document their work throughout the entire course of a research project, from conception of an idea to final publication and beyond.

The specific information to be included in an analysis plan will depend on the type of study. We expect to engage in an open and constructive dialogue with each research grantee and consultant as to the level of detail needed for each preregistration. Researchers should feel free to use the following list as a template for preregistration and even for preparing a research proposal. If they would prefer, they are welcome to share the information in narrative form. In either case, the researchers and LJAF personnel may have further conversations to flesh out particular details relevant to a given project.[2] To a large extent, many of the items below could be put into a set of "standard operating procedures" for all projects conducted by a researcher. Moreover, it may not be possible to specify all of the

[1] Researchers may also wish to preregister via ClinicalTrials.gov, SocialScienceRegistry.org, or others.

[2] When a study is written up in a final report, the CONSORT 2010 criteria should be followed as closely as possible, including the flow chart of participants. For a general explanation of CONSORT 2010, see here. For an explanation of how the CONSORT criteria might apply to the social sciences, see Boutron, John, and Torgerson (2009).

details below at the outset of a given project; in that case, it may make sense to prespecify as much as is feasible at the outset, and then update the registration when new details come to light.

1. **Study title**
2. **Principal investigator** (with name, affiliation, and Open Researcher and Contributor ID (ORCID) if available)
3. **Co-investigators** (with names, affiliations, and ORCIDs if available)
4. **Geographic location of study**
5. **Keyword(s)**
6. **Abstract** (300 word maximum)
7. **Projected timeline for the study** (start date and end date of funding; start date and end of intervention and/or data collection; and any other significant event).
8. **Other funding:** Identify any other sources of funding.
9. **Institutional Review Board (IRB)**: Identify details about IRB approval (date of approval, identity, and address of IRB).
10. **Hypotheses/research questions**: Describe the specific hypothesis or hypotheses that will be tested in the study.
11. **Outcomes**: Define your study's primary outcome(s) as completely as possible, including how and when they will be assessed. Do the same for any secondary outcomes.
12. **Sample size and statistical power:** State your study's sample size and provide details on statistical power calculations. If any clustering is involved, please be sure to account for the intra-cluster correlation coefficient, and describe how that coefficient was determined. We prefer that you begin by defining the minimum effect size that would make any real-world difference. Then, assuming at least 80 percent power and an alpha level of 0.05, determine the sample size. Please provide a code script (*e.g.,* in Stata or R) used to determine power or a detailed description of the specific module and design parameters used to calculate power in other programs (*e.g.,* Optimal Design or PowerUP!). Monte Carlo simulations are fine (indeed preferable) for many study designs that may not have a clean analytical solution.
13. **Data sources:** Describe the source(s) of data in detail.
    a. For studies involving existing datasets, provide or link to a description of those datasets, including where they are available, what years they cover, and the settings/locations where the data were collected. If possible, link to the datasets and a codebook and/or data dictionary.
    b. For studies involving the collection of new data, describe the settings/locations where data will be collected, the time period(s) for recruitment, the time period(s) for data collection, the time period(s) for follow-up, the specifics on all data that you plan to collect, and the procedures that will be used to ensure data provenance and quality. Plan ahead to create a codebook and/or data dictionary that would be usable by outside researchers.

14. **Intervention/policy:** Describe the intervention or policy being tested, in as much detail as is feasible.
    a. For randomized controlled trials (RCTs), please describe the control condition(s). For example, if there will be one control group that is "business as usual," describe what that looks like. If you will add an additional "active control" group, describe that active control condition and how it will accurately estimate a placebo/expectancy effect without contamination by an actual treatment effect.
15. **RCT-specific considerations**: If the study is an RCT, please provide the following:
    a. Describe the level of randomization and the level at which the outcome data will be collected.[3]
    b. Describe any further details about the randomization and design. This might include allocation ratio and probability of assignment to each condition, use of factorial designs, blocking or stratification, matched pair randomization, etc.
    c. Describe who, if anyone, will be blinded after random assignment (such as participants, providers, anyone assessing outcomes, or the research team themselves), and how.
    d. Describe any inclusion or exclusion criteria.
16. **Statistical methods:** Describe the statistical methods that are intended to be used (the following categories are not mutually exclusive).
    a. For RCTs:
        i. Plan to use an intent-to-treat analysis as the primary analysis, but indicate whether you plan to use randomization as an instrument to get the local average treatment effect, or perform some other analysis that differs from intent-to-treat.
        ii. If you plan to add control variables to increase precision of the estimates, or to account for stratification, blocking, or pair matching, indicate (if possible) what those will be.
        iii. Indicate whether you plan to perform any interim analysis of the outcomes before the trial is concluded, and if so, exactly when that will occur and based on what criteria.[4]
        iv. Indicate whether you will use permutation/randomization tests.[5]
    b. For regression discontinuity:

---

[3] For the latter, take Cameron and Miller's paper into account, where relevant. See http://cameron.econ.ucdavis.edu/research/Cameron_Miller_JHR_2015_February.pdf.

[4] We are primarily concerned about early peeking at the results in a way that might lead to changing the sample size by stopping earlier or later. It is sometimes necessary to change the timeline and sample for reasons of ethics, safety, or feasibility, but we want to be sure that this process is not unduly biased by excitement or disappointment over the impact estimates.

[5] See Alwyn Young's working paper: http://economics.mit.edu/files/11362.

i. Describe the running variable, where the cutoff or threshold occurs, and how you will choose the appropriate bandwidth.
ii. Describe your planned model in as much detail as possible.

c. For fixed effects studies:
i. Describe the variables that will be used as fixed effects, and how they will control for the relevant endogeneous (but time-invariant) factors.
ii. Consider using random effects instead, using the Mundlak (1978) correction if necessary.[6]

d. For difference-in-differences, comparative interrupted time series, synthetic control, etc.:
i. Describe how the control group will be chosen.
ii. Describe how many years of data will be collected from before and after the intervention/policy in question.
iii. Indicate whether you plan to use placebo tests (*i.e.,* different years, different sets of controls).

e. For any study involving matching of any kind:
i. Identify how you will choose the type of matching method (propensity score, coarsened exact matching, etc.).
ii. Identify the covariates to be used in selection.
iii. Identify the distance measure, and any blocks or stratification.

f. For machine learning:
i. Indicate the algorithm(s) to be used (including other pertinent details, such as kernel functions for support vector machines).
ii. Indicate techniques to prevent overfitting (cross-validation, data splitting, differential privacy, etc.).
iii. Indicate criteria used to judge success (AIC, BIC, error or loss functions, confusion tables, etc.).

17. **Additional statistical considerations:** How will the research team handle the following, if applicable?
a. Missing data, attrition, and censoring
b. Multiple comparisons adjustments[7]
c. Robustness checks and sensitivity analyses[8]

---

[6] See http://people.stern.nyu.edu/wgreene/Econometrics/Mundlak-1978.pdf and http://www.stat.columbia.edu/~gelman/research/unpublished/Bafumi_Gelman_Midwest06.pdf.

[7] Benjamini-Hochberg and the like are probably better than Bonferroni, given the risk of Type II error with the latter. The What Works Clearinghouse agrees (see Appendix D here).

[8] We appreciate techniques such as Patel et al.'s "vibration of effects" analysis (see http://www.sciencedirect.com/science/article/pii/S0895435615002772) or Simonsohn et al.'s "specification curve" analysis (see https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2694998).

II. Openness and Sharing

Unless prohibited from doing so, researchers should make the materials on the OSF page—datasets, code, pre-registered documents, etc.—public. Although these Guidelines refer to timeframes for designating items as "public," sooner is always better. We also want the results of *all* studies to be made public, including results that are null, negative, or messy and hard to interpret.

A. Computer Code

Researchers should already, as a matter of course, produce well-annotated scripts to clean and analyze data. The final version of these scripts should be uploaded to OSF or a similar site such as Github, and made publicly available. Ideally, the final code scripts should enable another researcher to take the original raw dataset(s), clean and merge them as was originally done, and re-run the original analysis.

We encourage researchers to use software that creates dynamic documents using literate programming, so that anyone else can easily duplicate analyses. We also recommend Gentzkow and Shapiro's paper "Code and Data for the Social Sciences: A Practitioner's Guide," which explains how many of the coding techniques often used by social scientists could be improved.

B. Datasets

We appreciate that many datasets are private and confidential, either because of federal law, state law, local law, IRB requirements, and/or agreements with the jurisdiction or agency that houses the data. We further understand that confidentiality obligations may extend to broader characteristics of the data, such as the geographic area or district from which the data was collected. In many of the research studies that LJAF funds, datasets may be subject to heightened confidentiality requirements (such as access being limited to specific individuals identified in advance or to computers that are not connected to the Internet, and requirements to destroy the dataset after use), and we will not ask researchers to take any actions with respect to datasets that would violate those requirements.

Yet, to the extent datasets are not subject to confidentiality requirements, we believe that datasets should be shared as freely as possible. At the most basic level, even the most well-intentioned scholars can make mistakes that would never be revealed unless someone else could double-check the code against the actual dataset. Data sharing also enables scholars to check others' work for sensitivity to the assumptions or model, as well as to extend it via further analyses.

Our general requirements regarding datasets are as follows:

1. When no restriction is to the contrary, datasets should be saved on external sites such as trusted digital repositories, because datasets are often lost or misplaced when saved only on individual computers. This can be done at OSF in the version-controlled system; OSF is also

connected to numerous other services, such as DataVerse, Amazon hosting, Dropbox, and more. Datasets saved on OSF are not public unless you take the separate step of designating them as such.[9]

2. To the greatest extent permissible, the dataset or some subset of it (*e.g.*, after removing personally identifiable information) must be made available for review by either the public or independent researchers at the time of the publication date of the initial study or by the end of the grant term, whichever is earlier.[10]
3. Include a codebook or data dictionary that would enable other researchers to understand the dataset and how it is structured.

In some cases, there will be no sharing requirement at all: the dataset in question simply cannot be released to anyone else under any condition, due to the highly confidential nature of the data or other statutory or regulatory constraints. In other cases, this might mean making the dataset available only on a highly restricted basis to a single third-party researcher who signs a non-disclosure agreement. Yet in other cases, this might mean removing all personally identifiable information and using anonymizing techniques on the dataset before posting it. This issue will be resolved on a case-by-case basis and memorialized in the relevant LJAF grant or consulting agreement.[11] In the event that the dataset-sharing requirement does apply, researchers should follow the guidelines in Inter-university Consortium for Political and Social Research's *Guide to Social Science Data Preparation and Archiving*, to the extent it is relevant.

C. Results of the Research

After a research project is completed, the grantee or consultant should ensure that the OSF webpage is updated to include the findings either in the form of a written report or a link to a publication or preprint elsewhere. These findings must be freely available in some form (not solely as a publication that is behind a paywall). The OSF webpage will thus provide a comprehensive overview of an entire research project from start to finish. Moreover, in the event that a research project does not lead to a peer-reviewed publication, posting the results at OSF serves a valuable informative purpose.

---

[9] Please note that OSF currently does not support encryption on the back end; therefore, if it is important that a specific dataset be encrypted so as to protect it even in the case of a server attack, the researcher would have to encrypt it. The same may be true of alternate repositories.

[10] We will consider on a case-by-case basis whether this guideline applies to the raw data or to a dataset that has been cleaned up for analytical purposes. As a general principle, the dataset should be available in a form such that another researcher could redo any transformations, rescalings, etc.

[11] There can be complex issues about how to make certain types of data available to other researchers (*e.g.*, biological samples for medical research). We will handle those on a case-by-case basis as well.