Plan Overview

A Data Management Plan created using DMPonline

Title: Copy of Who Benefits from Development Programs?

Creator: Jasmin Fliegner

Affiliation: University of Manchester

Funder: Economic and Social Research Council (ESRC)

Template: ESRC Template

Project abstract:

Each year governments, aid agencies, and NGOs, implement an array of development programs. While these programs take many forms including cash transfers, training, health interventions, and credit, they usually share twin goals: (i) increasing living standards of the poorest individuals in society, and (ii) reducing inequality (two of the UN Sustainable Development Goals). Our project asks, in a general sense, whether those goals are being accomplished.

We distil the question into an analysis of heterogeneous program impacts. Heuristically, when a program is launched, do its benefits tend to flow to the poorer potential beneficiaries, improving their standards of living and closing the gap to the rich? Or will we find the opposite, that it is the rich who benefit the most, widening existing disparities? Answering these questions is important for better understanding existing programs and designing new ones.

New work is needed on this question, because existing studies are mostly designed to analyse average impacts of individual programs. This means they typically do not have sufficient sample size to precisely estimate distributional impacts. The few evaluations that do can only tell us about specific settings, and reveal little about whether the twin goals are being achieved in general.

We solve the power problem by combining data from many randomized evaluations conducted over the past 20 years, massively increasing our effective sample size. By aggregating evidence across many trials, we can draw general conclusions about development programming as a whole. On average have the programs evaluated over the last 20 years achieved the twin goals, or just one, or perhaps neither?

Our project has three parts. First, we will collect datasets from a large sample of previously-conducted Randomized Controlled Trials that evaluate development programs. To do this we will exhaustively review the public repositories of Innovations for Poverty Action and the Abdul Latif Jameel Poverty Action Lab. We choose these organizations because they have wide coverage of contexts, authors, topics, and countries. We have already collected part of this data for a previous project, which demonstrates that large-scale reanalysis of the type we propose here is feasible.

Second, we develop a systematic approach to assessing whether the twin goals are met in each program individually. We divide (potential) beneficiaries into two groups, the “better off” and the “worse off” based on observable characteristics. Then, for each program in our sample we ask 1) which group was more likely to take up an offer of the program? 2) conditional on take-up, which group benefited more? And 3) overall (combining 1 and 2) did benefits accrue more to the better- or the worse-off? As noted above, this type of analysis is

often underpowered.

Third, we will aggregate the evidence. We transform the data into a standardised form so that findings from different studies can be compared on a level playing field. Then, we use tools from meta-analysis to assess whether, *on average across all programs,* it is the better- or the worse-off who benefit the most. These meta-analytic techniques will allow us to reap the power rewards of the hundreds of thousands of data points available across the many different interventions in our collection and also allow us to explore which types of programs are able to achieve the twin goals.

ID: 160428

Start date: 01-09-2024

End date: 30-08-2026

Last modified: 03-10-2024

Copyright information:

Copy of Who Benefits from Development Programs?

Assessment of existing data

Provide an explanation of the existing data sources that will be used by the research project, with references

The objective of this project is to answer the question “who benefits from development programs”. In order to carry out this research, we will be (1.a) handling and (1.b) transforming previously published de-identified datasets of randomized controlled trials (RCTs). This includes collecting metadata for each RCT dataset and writing estimation code. Based on those datasets, we will produce the main outputs: a replication package that contains (2) a newly created dataset with our estimation results which will the basis for our analysis and corresponding (3) analysis replication code and (4) a research paper that describes our research findings. No new primary research data will be collected throughout the tenure of this grant.

(1.a) Previously published datasets
Our sample of studies are drawn from the Datahub for Field Experiments in Economics and Public Policy, specifically from Innovations for Poverty Action (IPA) and the Abdul Latif Jameel Poverty Action Lab (J-PAL). Each contains tabular type data from previously-conducted and published RCTs conducted by affiliates of the organization, over 200 datasets in total (with some overlap). We expect one-third to one-half will be usable for our method. The datasets come in different formats such as tex-files, Stata-files or Excel files.

J-PAL and IPA have high standards for data publication, follow ethics requirements and assist the dataset owners during the publication process to protect human subjects. They specifically check for the risk of personal identifying information (PII) during that process. Therefore, we have high confidence that the data meets ethics standards, is already free of PII and therefore does not bring risk of publicising PII. The default license for data published on the J-PAL dataverse and to our understanding also on the IPA dataverse is CC0 1.0 Universal, which allows users to "copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission." Accordingly, we are allowed to transform and distribute the datasets.

Provide an analysis of the gaps identified between the currently available and required data for the research

We are relying extensively on currently available data. In order to answer our research question as proposed in this application, we are required to transform the previously published datasets and create an aggregate results dataset based on our estimation strategy outlined in the proposal. It is also important to treat each original dataset in the same way to achieve comparability of results and in order to aggregate them. To the best of our current knowledge, this question has not been addressed before the way we do and no currently existing dataset contains this aggregate information. Therefore, we are required to create the outputs mentioned in section 1 - namely a new results dataset, the research paper and the corresponding estimation code in R.

Information on new data

Provide information on the data that will be produced or accessed by the research project

(1.b) Derived Previously Published Datasets
Based on the previously published datasets from (1.a), we will create for each study a transformed dataset. If possible, we will use automated procedures to read in metadata but information about the study might have to be collected manually by reading the accompanying information. We will use the software R to produce readable CSV-files (or a similar international standard to save tabular data) containing the metadata and transformed databases for each study. We may realise that a dataset for which we have started collecting metadata cannot be grinded into our approach. We will then abandon it during the process and not transform it.

(2) Results Dataset
Based on the already published datasets, we will create a new tabular dataset with aggregate results and metadata from the original RCTs. Since this dataset only contains aggregate information from the originally published datasets, we even further limit the risk of handling data containing PII. The dataset will be a CSV-file (or a similar international standard to save tabular data). We will use the software R to estimate and analyse those results.
(3) Code
We will use the software R to read the datasets, produce our results and analyse them.

(4) Research paper
We will produce a research paper that describes our findings in a PDF document.

We consider that all of these data are of long-term value and can be shared with exception of the abandoned previously published datasets (see *(1.b)* above). The relevant metadata and derived datasets from the studies that we actually use in the analysis for the paper will be shared along with the Results dataset and code. A readme file in an appropriate format such as PDF will be included in the replication package to guide potential users.

Quality assurance of data

Describe the procedures for quality assurance that will be carried out on the data collected at the time of data collection, data entry, digitisation and data checking.

As described above, we consider the previously published datasets of excellent quality due to the high standards at J-PAL and IPA for data curation. In order to create the derived data, we will use a standardized estimation code in R to ensure that each dataset is treated the same way. In order to capture the metadata from the originally published dataset, we will provide detailed step-wise instructions for anyone working on the creation of those including the research assistant. Where applicable, we will also apply data validation techniques such as using multiple choice answer options. We use peer review to ensure quality. The code is jointly monitored, written and reviewed by the team. The research assistant and metadata entries are supervised by a team member.

Backup and security of data

Describe the data security and backup procedures you will adopt to ensure the data and metadata are securely stored during the lifetime of the project.

Due to the non sensitive nature of our used, transformed and created data, we do not expect to implement additional security measures. Should we unexpectedly find PII in the previously published datasets, we will inform the repository and the data owner, subsequently not use this version of the data and destroy our copies of the data until the data owner republish a new, fully de-identified version of the data.

For short-term storage during the tenure of the grant, we will use a combination of local and secure cloud computing resources as appropriate for the project, and keep a copy on institutional storage at Queen Mary (for backup and redundancy). We will have access to QMUL's OneDrive for Business (ODfB) and SharePoint to store data for collaboration. In addition, we will have access to the QMUL HPC research storage platform. We will ensure at least two copies of data exist on two separate storage platforms. According to QMUL IT policies, all data stored on these two platforms are subject to routine backup and there are processes in place to ensure that the data is recoverable with minimal backup in case of system failure.

During the drafting process, the research paper describing our findings will be stored as .tex document on overleaf.

Management and curation of data

Outline your plans for preparing, organising and documenting data.

We seek to prepare our data for future sharing and re-use as described further in section 10. We will ensure that all data are well organized and that the replication package contains a detailed readme file to facilitate re-use and increase transparency.

Difficulties in data sharing and measures to overcome these

Identify any potential obstacles to sharing your data, explain which and the possible measures you can apply to overcome these.

Due to the non sensitive nature of our used, transformed and created data, we do not expect any major risks to data sharing. We address the data specific risks (risk of losing data due to IT problems, delay in data sharing and finding PII in the previously published datasets) in the relevant sections and explain how we aim at dealing with these situations.

Consent, anonymisation and strategies to enable further re-use of data

Make explicit mention of the planned procedures to handle consent for data sharing for

data obtained from human participants, and/or how to anonymise data, to make sure that data can be made available and accessible for future scientific research.

Due to the non sensitive nature of the data as described in section 1, no additional procedures need to be put in place to be able to share the data.

Copyright and intellectual property ownership

State who will own the copyright and IPR of any new data that you will generate.

We will share the derived datasets and the replication package data under the license CC0 1.0. For the derived datasets, we will always make sure we comply with the license of the original data publication. For example, should we unexpectedly find a dataset that is under another license than CC0 1.0 that does not allow us to distribute the data, but use and modify them, we will only publish the replication code but not the transformed data.
Intellectual property remains with the authors.

For the paper, once accepted for publication, we would make the paper open access according to the journal rules. Intellectual property remains with the authors.

Responsibilities

Outline responsibilities for data management within research teams at all partner institutions

Our team is committed to protecting the confidentiality and integrity of those included in this research and seek to make our data open access and encourage secondary analysis. All team members and every research assistant working on this project will follow the data management plan outlined below. This will be ensured during regular meetings and research assistant supervision.

Preparation of data for sharing and archiving

Are the plans for preparing and documenting data for sharing and archiving with the UK Data Service appropriate?

(1.b) Derived Previously Published Datasets
For long-term storage, we will (re-)publish the derived datasets at the UK Data Service Repository under the license CC0 1.0, acknowledging the data owner and original data repository using appropriate citation (e.g. DOIs). We will transform the data in a suitable long storage format as required by the repository. We aim at completing this three months after the end of the award, except where this conflicts with the policies of the journal our paper will be published in. In such a case we would ask for an embargo in line with principle 5 of the ESRC Research Data Policy.

(2) Results Dataset

We will follow the same procedures as for the previously published datasets (including license and timeline). Since this dataset only contains aggregate information from the originally published datasets, we even further limit the risk of publishing data containing PII.

For our prior project Bernard et al. (2023), we created an interactive website to disseminate our results. We plan to create a parallel site for this project as well so that users, policy makers, academics or even teachers, can investigate the results themselves, or get a quick and user-friendly overview of the main results using summary statistics.

(3) Code

The code will be stored via Github and made openly available for re-use as part of the replication package either on the Github project website or alongside the results dataset on the data repository.

(4) Research paper

We will share the PDF working paper with our research community as a working paper. We then have the intention to submit this paper for publication in a high-impact economics journal ensuring that all data used and the funder are appropriately acknowledged and point to the replication package. Once accepted, we would make the paper open access according to the journal rules.

Is there evidence that data will be well documented during research to provide highquality contextual information and/or structured metadata for secondary users?

For documentation and findability, we make sure to use informative metadata to describe all outputs and create a DOI for the replication package in accordance with the repository to comply with the FAIR principles.