Data for Development Impact guide Proposal

This proposal is for a short practical publication titled "Data for Development Impact: The DIME Analytics Resource Guide". The guide's primary purpose is to be a workflow guide for DEC & DIME research assistants (RAs), and should serve as a desk reference for people of all experience levels who use data as part of their development workflow. It will also have an online version, similar to the structure of Kieran Healy's Data Visualization.

Data for Development Impact will serve as a companion to the DIME Wiki, and heavily reference and link to appropriate Wiki pages. The guide's original content primarily comes from the addition of specific *workflows*: how-tos that illustrate specific tasks with specific software combinations and commands. Where the Wiki seeks to explain *why* things are done in a manner that improves the readers' understanding of topics, the Guide provides a linear, narrative structure to the Wiki content, leading readers through specific tasks.

In this sense, the Wiki functions much like a "user manual": once someone knows what they want to know, they can explore in a rich and self-structured way the full available information on that topic. By contrast, this publication is a "user guide": it offers a very quick reference introduction to the topics that the user needs to know, key concepts they may need to research further, and specific instructions for getting through the simpelst version of the specific task at hand. In further contrast to a *theoretical* toolkit like Using Randomization in Development Economics Research, it is envisioned as a *practical* toolkit – each short section should allow the reader to implement the most basic version of the task using code snippets and available or simulated data.

The guide will be structured according to the four thematic sections of the DIME Wiki (pictured). However, the content will be distinct: the guide should serve as a readable and practical guide to core tasks of research design, data collection, data analysis, and publication. It will provide concrete guides to tools and processes used at DIME, and will also incorporate `code snippets` as a handy reference. It will have a practical guide to code collaboration and version control with

Git; and include a guide to the importance of quality code and integration with the `iefolder` file structure.

It will require updating significant portions of the Wiki to incorporate materials from the DIME Onboarding and Continuing Education, the Manage Successful Impact Evaluations training, and materials from the DIME Analytics GitHub repositories. It would be therefore desirable to hire a full-time STC for curating, cataloguing, and updating these resources as necessary to complete the guide.

For example, the first part of the "Sampling and Power Calculations" Wiki page reads:

> Sample Size
>
> Calculations are a statistical tool to help determine Sample Size. This is important, a sample that is too small means that you will not be able to detect a statistically significant effect, and a sample size that is too large can be a waste of limited resources. You can estimate either sample size or minimum detectable effect. Which you should estimate depends on the research design and constraints of a specific impact evaluation. The types of questions you can answer through power calculations include:
>
> - Given that I want to be able to statistically distinguish program impact of a 10% change in my outcome of interest, what is the minimum sample size needed?
> - Given that I only have budget to sample 1,000 households, what is the minimum effect size that I will be able to distinguish from a null effect? (this is known as Minimum Detectable Effect)
>
> Power calculations should be done at Impact Evaluation Design stage. They are mostly typically done using Stata or Optimal Design (See Power Calculations in Optimal Design, Power Calculations in Stata). Power calculations can be used to determine either sample size (using standard assumption of 80% power) or power (if sample size is constrained).

The linked page, "Power Calculations in Stata", has a long list of options and comparisons between power calculation options. By contrast, this guide's section on "Sampling and Power Calculations" would:

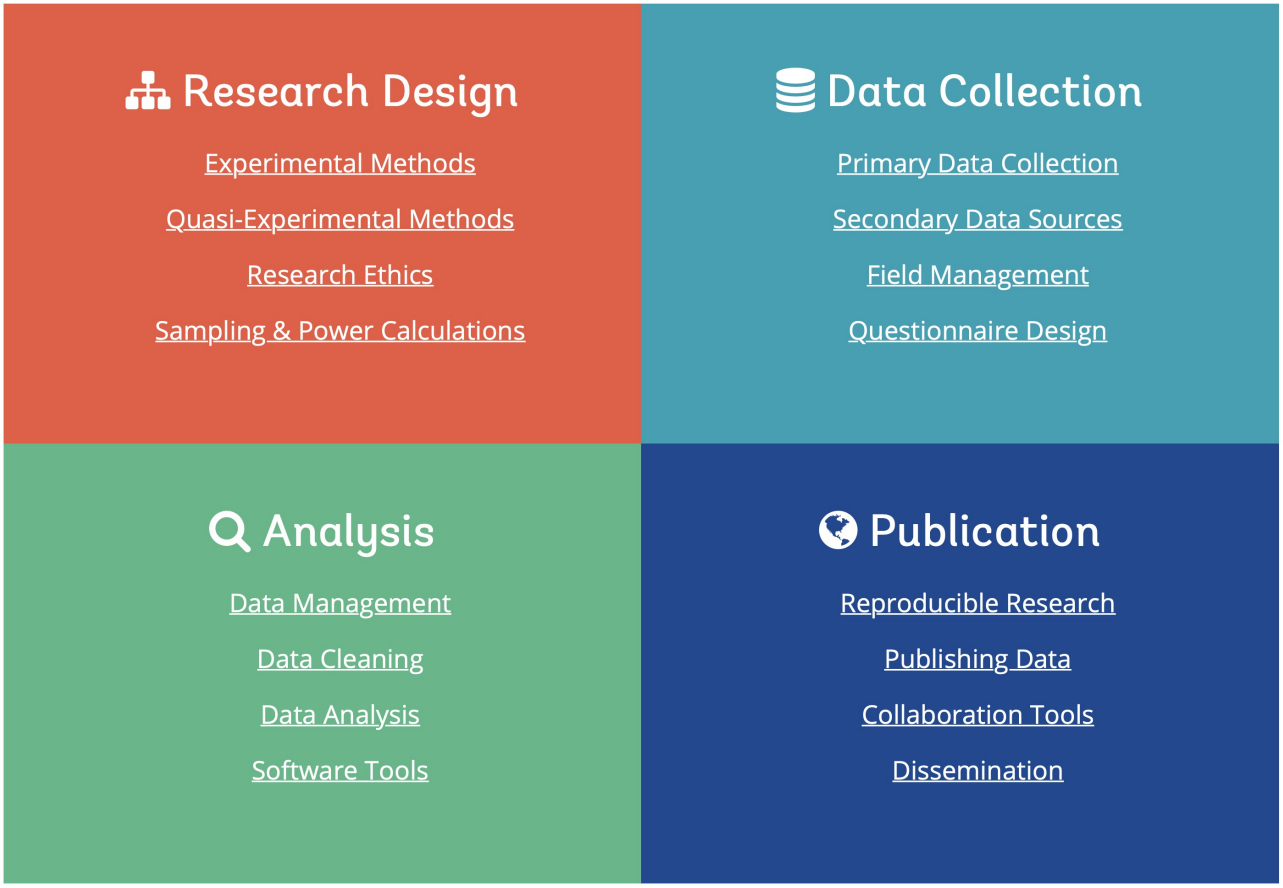
1. Motivate power calculations with a short paragraph, referencing (and linking) one or two preeminent publications on the topic ("Why most published resesarch findings are false").
2. Use Stata to (a) sample and (b) randomize from a universe dataset.
3. Use a simulation approach to repeat (2) many times with simulated outcomes, illustrating the dispersion of the regression coefficients due to each part of the process with fixed

parameters.
4. Show that the power of an experimental design is a function of sampling and randomization parameters by varying those parameters and recording power.
5. Show that the power of an experimental design further varies with the hypothesized effect size, and provide a final simulation doing so.

The total written text for this would be only a few short pages, with the primary resource being commented code and illustrations. These will be linked in a GitHub repository or other similar public resource where they can be copied and run.

Proposed Table of Contents

Four thematic sections in the DIME Wiki



Introduction: Data for Development Impact

This section motivates the guide. Drawing from sources like Best Practices for Scientific

Computing and Code and Data for the Social Sciences, the introduction describes what Data Science has to bring to development economics. Broadly, it is our idea that technical computing skills and paradigms are underinvested in, compared to mathematics and topical expertise in economics research. This means that RAs who want to do their job efficiently currently have very little in the way of mentorship or guides to their work and the conventions, standards, and best practices that are fast becoming a necessity for impact evaluation projects.

Research Design

Experimental methods for impact evaluation

This section reviews the most common experimental methods for impact evaluations, drawing on Impact Evaluation in Practice, Causal Inference: The Mixtape, and The Econometrics of Randomized Experiments. It covers the design of RCTs for difference-in-difference and regression discontinuity designs, and points to randomization code for treatment assignments.

Quasi-experimental methods

This section covers instrumental variables, matching estimators, and synthetic control models.

Sampling and power calculations

Building off the twin ideas that "a regression is a random variable" and that "a randomization is a random variable", this section introduces the ideas of sampling noise and randomization noise as justifications for standard errors on effect size estimates. It suggests simulations based on both, providing templates for power, minimum detectable effects, and sample size calculations based on asymptotic inference and on randomization inference. It provides guides to reproducibly sampling and randomizing using `sample`.

Data Collection with `iefieldkit`

Primary data collection with SurveyCTO

This section details the various ways in which a product like SurveyCTO can be used for data entry, including both field use of tablets as well as Web-based entry of paper forms. It details encryption of surveys and use of the Sync application to download data.

Questionnaire design with SurveyCTO

This section provides a basic overview of key SurveyCTO form options and conventions. It emphasizes conventions for section and variable naming, use of language labels to improve data readability, and quality controls such as bounds and logic checks that can be embedded directly into forms.

Field management and quality assurance

This section details the data quality checks which are commonly used throughout fieldwork. This includes enumerator checks, high-frequency checks, tests for outliers, and schedule/sample completion reports. It emphasizes the RA's responsibility to spot, rather than solve, problems at this stage, and provides methods for organizing paper trails for issues and corrections in data.

Managing primary and secondary data sources

This section focuses on data storage, including cloud storage, sharing, and backup.

Data Analysis with `ietoolkit`

Data and code management

This section outlines the function of the `iefolder` command and the folder structure that it creates. It describes what each folder should be used for and outlines the overall workflow in terms of this folder structure, from raw PII data to final outputs. This section outlines core principles from computer science for impact evaluation practitioners. It focuses on the ideas of modularity, generalizability, and anti-repetition. It suggests adopting a flexible and extensible text editor, using descriptive naming conventions for data and code work, and planning to routinize repetitive tasks.

This also section emphasizes that modern coding is typically collaborative, with multiple people usually needing to simultaneous work on or access code, and be sure of its functionality and location. This section first outlines some basic ideas for code organization and readability, then provides a short technical guide to setting up and working on a project in Git/GitHub using the Git Flow workflow model and emphasizing that, like `iefolder`, a good organization model is an mental tool rather than a technical solution.

Data cleaning

This section takes the reader through the process of cleaning and deidentifying raw survey data, beginning with a spreadsheet download from a source such as SurveyCTO. It emphasizes

creation of a Raw-Deidentified dataset, followed by Intermediate and Constructed datasets, which can be kept in a shared folder for collaborative use. It focuses on conventions for variable naming at each stage and provides syntax and examples for common commands such as `reshape` and `merge`, and suggests that Constructed data should regularly have as many purpose-built datasets as needed.

Data analysis

This section reviews the common estimators for each of the methods outlined in the Experimenal Design section, with references to code packages that are provided in `ietoolkit` and elsewhere for implementing them. It emphasizes that data cleaning and construction should not be done in analysis dofiles, and suggests a modular approach to outputs that strengthens the link between the creation code and the product. It provides conventions and checklists for exporting results as tables, and for organizing and naming dofiles and outputs.

Data visualization

This section presents some common styles of data visualization, key design elements, and a short guide to editing key graph elements in Stata.

Publication

Doing research reproducibly

This section emphasizes that analysis code should be written with the idea that it is an output of the research as much as the completed paper is. Analysis code should be written primarily for others to read and run, and as a methodological investment for re-use in the future. This section suggests routinization of analysis tasks that are done repeatedly, with the goal of producing extremely short core dofiles.

Publishing data and code for replication

This section provides tools and guides to commenting code, setting up public-release repositories on OSF or GitHub.

Collaboration tools for academic writing

This section is a guide to using LaTeX for preparing research papers, with collaboration either via GitHub or Overleaf. It provides a guide to bibliography management with BibTeX; sources for

document templates; checklists for outputs (page numbering, author affiliations, etc), and methods for converting TeX files into formats like .docx if needed.

Research ethics and requirements

This section covers ethical requirements for conducting research with human subjects, handling personally-identifying information, and transparency in research design. It includes links to pre-analysis plans, the NIH's Protecting Human Research Partipants course (old version) and the CITI program, and information about encryption and transfer of sensitive data.