Big Data for Development
Prof. Sam Asher

Final Paper Guidelines

Overview

The final paper is a 2200-2700 word piece of quantitative empirical research on a topic of your own choosing, using theory from the class to guide your analysis, econometrics to test your hypothesis, and your knowledge of the literature to situate your findings in a broader body of evidence.

In terms of topic, the choice is yours, beyond the fact that it must relate to the topics covered in this class. The purpose is for you to become an expert in a topic of your choosing and attempt to provide new evidence on that topic. This is a good chance to gain expertise and generate a writing sample in an area that you might want to work in in the future. If you are interested in a topic in growth and development economics that we will not be able to cover in this class (e.g. the environment), please send me an email with your idea and I will be happy to consider whether it is a growth/development economics topic. I only have three rules regarding the choice of topic: a) it must be about something related to growth/development economics, b) it must estimate a causal relationship, and c) you must use "big" data, defined broadly.

The purpose of this paper is to conduct original data analysis. So you'll need an explicit research question, an empirical strategy, and a dataset with which to test your hypotheses. If you don't have all three of these, you don't have the makings of a paper. You can start to converge towards a paper starting from any one of these. I've written papers that started with an interesting new dataset, where I then figured out the best questions that could be answered with it. Alternately, you can start with a topic of interest, think about where the literature is lacking, and then find the data with which to fill the gap. Your empirical strategy should include some regressions to test for the causal relationship between variables, but tables and figures can also include summary statistics, scatter plots, maps, etc.

That said, this is a single final paper in an undergrad class, not a PhD dissertation chapter. While it might just eventually turn into something you want to publish, that's not the objective here. You have limited time to read books/papers, come up with a question, find the appropriate data, conduct the analysis, write up your results, and then rewrite to produce your final paper by the due date. So my expectation is not by any means that you generate research that breaks entirely new ground or fills an important hole in the literature. It's fine to reproduce an existing paper in a new context or end up with insignificant results (you will not be graded on what you find but on your execution and writing), and I expect that there will be some unresolved problems in your empirical strategy.

Because this is a big data class, you will likely be merging multiple datasets together (e.g. a household survey + gridded weather data), but my recommendation is to keep complexity to a minimum. I'm most knowledgeable and able to provide data on India, but I'm happy to advise on data from other countries as well. I highly recommend that you speak with the TA and me about the paper throughout the term. I'd be very happy to read a brilliant paper that received little guidance from me, but papers are always better when the ideas have gone through multiple iterations and have received careful feedback early in the process.

Writing an original paper is hard, in part because there are infinite potential empirical tests, ways to write up your findings, etc. I find it helpful to base the outline of my paper on a published paper that I think is particularly well done and is of a similar type to mine. For example, if you are writing a regression discontinuity paper, you may want to base it on my paper "Rural Roads and Local Economic Development" (AER 2020).

Additional guidance**:**

- Please use the following paper structure, although you should feel free to add sections that you consider necessary for the reader to understand what you have done, why, and what to conclude from your analysis: (i) introduction, (ii) literature review, (iii) data, (iv) empirical strategy, (v) results, (vi) conclusion, (vii) references. Put your tables and figures in the paper where you discuss them. The cover page should have the title of the paper, the date, and a 100 word abstract summarizing the paper for someone who isn't going to read any further (good practice for figuring out how to succinctly summarize what you've done, what you find, and why it matters).
- Results should be presented in clean tables and figures that are placed in the body of the paper where they are described rather than at the end. (This makes it easier for me to read.) Do not just copy and paste Stata/R/etc code and output into the paper! Look at how tables and figures are structured in the economics papers we read to understand how to report results. While there is no hard and fast rule about how many figures and tables you have, I would be surprised if a thorough paper had fewer than three of each, and likely more.
- Every table and figure in your paper should have a note explaining exactly what it does. See table/figure notes in published papers to understand what is needed.
- Citations: name and date in line and then full citation in the references at the end. I strongly recommend using Mendeley, Zotero or a similar references software package

that allows you to easily manage your references and automatically generates your bibliography.

- Make clear why we should care! What are the theoretical ramifications of your findings? How should policymakers think differently armed with your results? The point of doing research is to change our understanding of the world. You have to tell the reader how they should think differently about growth and development in light of your findings.
- The word count applies to words in the main body of the paper, not to tables/figures, table/figure notes, references, etc.

Due: January 12, 2026, 9 am (uploaded through the course website)

Data Guidance

For your final paper, you need to find “big” data appropriate to your research question and then analyze it. I get a lot of questions from students about the right data to use and how to find it. This document lays out some broad guidance for how to think about identifying and obtaining the data for your final paper.

First, some broad principles:

- We have taken a very broad definition of “big” data in this class, which is to say any large datasets that require handling in ways that are different from “regular” data like household surveys, country-year data on economic performance, etc. Examples are geospatial data (e.g. from satellites), textual data, data exhaust from government programs, digital trace data from cell phones, etc. You may want to merge these to rich household survey or other data in which you can measure your treatment of interest or perhaps your outcomes.
- The data must suit your research question. So if you’re interested in studying the effects of migration, your final analysis dataset needs to have variables on migration and on relevant outcomes (employment, wages, etc).
- Be opportunistic: it is likely that you will need to go through multiple iterations of research questions and datasets. Don’t get too attached to one or the other, or else you probably won’t be able to write a very good paper. You might start out interested in China, but realize that the data is actually best (or just available) in Nigeria.
- You can start with the data: some people start with very narrow research questions, look for data to answer that question, and then usually revise the question once they understand more about the strengths and weaknesses of that particular dataset. But others start with a broad topic, find rich data, and then figure out the best question(s)

that can be asked with it. I myself often find exciting new data and then think of the coolest questions I can ask with it.

- Read! Economic research does require creativity, but mostly it requires standing on the shoulders of giants. Reading papers in your area of interest will expose you to datasets you didn't know about, professors you can reach out to, empirical strategies that you can use in a different context, etc. Good papers are mostly original combinations of existing data, methods, hypotheses, etc.
- Remember that while you will be graded on the quality of your work, I am not expecting you to write a paper that is going straight into *Econometrica*. I know this assignment is hard given your research experience, coding skills, time limitations, etc. The goal is to produce a thoughtful piece of research, which means asking a good question, figuring out the right empirical strategy and data to test your hypothesis, and a thoughtful writeup of what you find and the limitations of your analysis.

On to the data – there are numerous ways to go about finding data for your paper. Here are some places to start:

- World Bank Microdata Library
    - LSMS: World Bank household surveys of LMICs around the world, often with multiple rounds and geocodes.
- Household microdata
    - Demographic and Health Surveys - USAID-funded, standardized surveys around the world (400 surveys across 90 countries, also available in standardized format through IPUMS). Many rounds have geocodes.
    - India Human Development Survey - 2 rounds (2005-6 and 2011-12) of 40k households in India (panel). I have these data in a clean form, so can share with you.
    - Middle East labor force surveys from the Economic Research Forum
    - Chinese Household Income Project Series (most recent round not there, but can be requested from CIID)
- Harvard Dataverse: people often post data from their studies there, especially from RCTs (but remember the big data requirement!)
- GIS Data
    - Global Human Settlement Data: so much gridded data on cities, land use, population, etc.
    - Night lights
    - Google Earth Engine
    - Facebook High Res Population Density Data
    - Climate Trace
    - Climate Data from NOAA
- Other household surveys with geospatial information
    - Afrobarometer

    - Unicef MICS
- SHRUG India Open Data Portal (from my lab Development Data Lab)
- Text as Data
    - Example: Train a classifier, e.g. on names
    - Political speeches, judicial decisions (but remember, you need to be working on a question that is broadly connected to development econ)
    - World Bank reports
    - UN resolutions
- Scraping Data – but warning, it can be hard!
    - News: Lexis/Nexis, BBC, or other news sources
    - Data from government websites