

Education Quality and Development Accounting: Data Documentation

Todd Schoellman*

December, 2015

Abstract

This supplemental file describes how to use the included Stata programs and data files to re-create the results, figures, and tables of the paper. It also provides the original Stata programs to estimate the returns to schooling of foreign-educated immigrants as done in the baseline analysis and all robustness checks. Finally, it documents the issues arising from the changes in Census data since the original empirical implementation, and shows how to re-create the original results given the currently available data. A Stata program is provided that performs this task for the baseline results of the paper.

*Address: Department of Economics, Arizona State University, Tempe, AZ 85287. E-mail: todd.schoellman@gmail.com

This supplement details the creation of all data, figures, tables, and regressions for the paper. There are six Stata do files and some of the data (but not all raw data) for re-creation of the results and robustness checks of the paper. The code is logically ordered, to be run as follows:

1. `returns_estimate_us.do`: Estimates returns to schooling of foreign-educated immigrants in the U.S. 1980–2000 Censuses.
2. `returns_estimate_nonus.do`: Estimates returns to schooling of foreign-educated immigrants in the Canadian 1981–2001 Censuses.
3. `returns_estimate_morg.do`: Estimates returns to schooling of foreign-educated immigrants in a pooled sample of the 1998–2002 NBER CPS Matched Outgoing Rotation Group samples.
4. `dataaggregate.do`: Formats and combines all point estimates of returns to schooling. Adds on other external data, such as the Barro/Lee schooling data. Produces two final data files, one of country-level observations (including returns to schooling) and a somewhat less important file with figures for the distribution of the population and wages across school levels. These two files are the inputs for `countryresults.do` and the final step of the paper.
5. `countryresults.do`: Takes the two external data files from `dataaggregate.do`. Uses them to produce estimates of η and cross-country variation in h . Compares variation in h to the literature. Produces tables and figures for the paper.
6. `production_function.do`: Estimates human capital using only assumptions about the aggregate production function as discussed in the brief note on my webpage.

Programs (1)–(3) produce the inputs to (4), which in turn produces the inputs to (5). Alternatively, (1) produces the inputs to (4), which produces the results in (6), only for development accounting.

To re-create only the aggregate figures, tables, etc., readers should take the two data files provided in this folder (`edqual_devacc_data.dta` and `edqual_devacc_data2.dta`) and run the file `countryresults.do`. Stata will store some temporary files along the way in a local subfolder `/temp`, and produce all figures and results in a second local subfolder `/output`. It also outputs the various estimates of η , of cross-country variation in h , and so on. This is logically the last step of the paper, but also the shortest. The rest of this file is dedicated

to explaining how to re-estimate returns to schooling of foreign-educated immigrants for interested readers.

Estimation of the returns to schooling of foreign-educated immigrants has been complicated by two changes in the data since the original empirical work was undertaken. First, the Census PUMS files that underlie the IPUMS data have changed. This was in response to a problem with the age recoding documented by Alexander, Davern, and Stevenson (2010). The Census Bureau provided a new age variable for the 2000 5% sample. Since age factors into sample selection and potential experience, these changes affect my sample composition and results. Second, the IPUMS system has changed its educational recoding from `educ99` to `educd`. `Educd` actually contains slightly more information than `educ99`. Again, since education factors into sample selection and acts as a regressor, these changes affect my sample composition and results.

Fortunately, the Census sample includes both the new and old age variables. Likewise, it is possible to collapse `educd` to recreate the information that was available in `educ99`. Hence, it is possible to re-create my original results using the current IPUMS data. I include a new (as of February, 2012) program, `returns_estimate_new.do`, that does precisely this for Table A1, the baseline estimates of the returns to schooling of foreign-educated immigrants. It is also natural to wonder whether the new data change the fundamental conclusions on the patterns of returns to schooling for immigrants from different countries. The program shows that this is not the case. Indeed, the estimated returns to schooling are nearly identical for the new and old data.

The raw census data files were downloaded from the Integrated Public Use Microdata Series, available online (Ruggles, Alexander, Genadek, Goeken, Schroeder, and Sobek 2010, Minnesota Population Center 2010). Readers who refer to their website will find that they can download the raw data files themselves. It is necessary to specify the samples, years, and variables. As of February 2012, users should specify the 2000 U.S. Census, 5 percent extract. Variables required are: `region`, `perwt`, `ageorig`, `age`, `sex`, `bpld`, `yrimmig`, `speakeng`, `educd`, `ind`, `classwkr`, `occsoc`, `wkswork1`, `uhrswork`, `incwage`, and `disabwrk`.

The remaining files only work exactly with the original Census extracts (available from the author upon request). However, the same basic concepts for converting age and education in `returns_estimate_new.do` will work for these files as well. Hence, I describe their contents and purpose for completeness.

The Stata programs (1)–(3), along with the appropriate raw data (not included), are sufficient to estimate all returns to schooling, including robustness checks. However, many of the raw data files are quite large, and the models estimated have a number of dummy and

interaction terms. Readers should be aware that the file requires Stata to allocate roughly 9gb to estimate the most complicated models. The baseline model and simpler variants can be estimated with somewhat less memory.

Each data file works with different raw data. The full list of Census extracts and variables needed is:

1. 2000 U.S. Census, 5 percent extract. Variables required: region, perwt, age, sex, bp1d, yrimmig, speakeng, educ99, ind, classwkr, occsoc, wkswork1, uhrswork, incwage, disabwrk.
2. 1990 U.S. Census, 5 percent extract. Variables required: region, perwt, age, sex, bp1d, yrimmig, speakeng, educ99, occ1990, ind, classwkr, wkswork1, uhrswork, incwage, disabwrk.
3. 1980 U.S. Census, 5 percent extract. Variables required: region, metaread, perwt, age, sex, bp1d, yrimmig, speakeng, higrade, occ1990, ind, classwkr, wkswork1, uhrswork, incwage, disabwrk.
4. 1981-2001 Canadian Census, 2.0, 3.0, and 2.7 percent extracts (pooled). Variables required: year, sample, serial, gq, provca, pernum, wtper, age, sex, marst, bplctry, yrimm, spkeng, educa, empstat, empstatd, classwk, hrswrk1, incwage, ca81a_wkweek, ca91a_wkweek, ca01a_wkswkp.

In addition, I estimate returns to schooling using the NBER's CPS MORG data.¹ I pool the 1998–2002 CPS data for this purpose. Finally, program (1) uses data on the licensure status of occupations. This data was collected from CareerOneStop (2010). Since this database is updated periodically, downloading this data again will yield different data than what I used. Mine are included in the subfolder licensure for replication purposes, as well as the necessary crosswalk to convert occupation schemes.

Programs (1)–(3) use Stata's `parmest` command, which writes the regression parameter estimates to external data files. The main function of the program `dataaggregate.do` is to read in all these poorly formatted parameter estimates and turn them into a more usable data file with country names, ISO codes, and so on.

In addition to this, the file `dataaggregate.do` combines together data from a number of other fairly standard sources which I do not include here. Schooling data comes from Barro and Lee (2001); internationally standardized achievement test scores data from

¹Available online at <http://www.nber.org/morg/annual/>.

Hanushek and Kimko (2000) and Hanushek and Woessmann (2009); returns to schooling from non-migrants from Banerjee and Duflo (2005); and literature results come from Hall and Jones (1999) and Hendricks (2002). PISA and TIMSS test scores (mentioned briefly) are taken from various reports in the case of PISA (Organization for Economic Development 2001, Organization for Economic Development 2003, Organization for Economic Development 2004). Test scores are taken from the results section of the website in the case of TIMSS (see <http://nces.ed.gov/timss/results.asp>). In each case I took the most recent math section score available for each country. These data are already included in the file `edqual_devacc_data.dta`, so it should not be necessary to return to the original sources for them.

`dataaggregate.do` produces two final data files, `edqual_devacc_data.dta` and `edqual_devacc_data2.dta`. These are included attached for interested readers. `countryresults.do` uses these two files to produce the final results, as detailed at the beginning of this file.

References

- ALEXANDER, J. T., M. DAVERN, AND B. STEVENSON (2010): “Inaccurate Age and Sex Data in the Census PUMS files: Evidence and Implications,” NBER Working Paper 15703.
- BANERJEE, A. V., AND E. DUFLO (2005): “Growth Theory Through the Lens of Development Economics,” in *Handbook of Economic Growth*, ed. by P. Aghion, and S. N. Durlauf, vol. 1A, chap. 7, pp. 473–554. Elsevier Science, North-Holland Publishers.
- BARRO, R. J., AND J.-W. LEE (2001): “International Data on Educational Attainment: Updates and Implications,” *Oxford Economic Papers*, 53(3), 541–563.
- CAREERONESTOP (2010): “Licensed Occupations,” Available online at http://www.acinet.org/ACINET/LICENSEDOCCUPATIONS/lois_keyword.asp?nodeid=16&by=keyword.
- HALL, R. E., AND C. I. JONES (1999): “Why Do Some Countries Produce So Much More Output Per Worker Than Others?,” *Quarterly Journal of Economics*, 114(1), 83–116.
- HANUSHEK, E. A., AND D. D. KIMKO (2000): “Schooling, Labor-Force Quality, and the Growth of Nations,” *American Economic Review*, 90(5), 1184–1208.
- HANUSHEK, E. A., AND L. WOESSMANN (2009): “Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation,” NBER Working Paper 14633.
- HENDRICKS, L. (2002): “How Important Is Human Capital for Development? Evidence from Immigrant Earnings,” *American Economic Review*, 92(1), 198–219.
- MINNESOTA POPULATION CENTER (2010): “Integrated Public Use Microdata Series International: Version 6. [Machine-readable database].,” Minneapolis: University of Minnesota., <http://www.ipums.org>.
- ORGANIZATION FOR ECONOMIC DEVELOPMENT (2001): “Knowledge and Skills for Life: First Results from the OECD Programme for International Student Assessment (PISA) 2000,” OECD Publishing.
- (2003): “Literacy Skills for the World of Tomorrow: Further Results from PISA 2000. Addendum with Data for Romania,” OECD Publishing.

——— (2004): “Learning for Tomorrow’s World: First Results From Pisa 2003,” OECD Publishing.

RUGGLES, S., J. T. ALEXANDER, K. GENADEK, R. GOEKEN, M. B. SCHROEDER, AND M. SOBEK (2010): “Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database],” Minneapolis: University of Minnesota.