# Accessing Administrative Data to Study High Skill Immigration

Kirk Doran

**Experience with Administrative Data Sets and Challenges to Working with Them:**

I have had extensive experience with two relevant administrative data sets. I will briefly describe them as well as how I obtained access to them.

(1) United States Citizenship and Immigration Services (USCIS) data on the Fiscal Year (FY) 2006 and 2007 H-1B visa lottery winners and losers.

For the FY 2006 and FY 2007 lotteries, the USCIS did record the full application information for both winners and losers of the lottery. In all other lottery years, the USCIS did not open the H-1B visa I-129 applications that lost the lottery, and did not record their mailing addresses either, but merely returned the unopened envelopes to sender. This is relevant because without lottery losers, researchers must resort to Department of Labor (DOL) Labor Certification Applications (LCAs), which contain many firms which decided not to file for I-129s in the end, thus biasing the results of any comparisons with lottery winners (all of whom did file for I-129s).

We obtained this data only after identifying an employee of USCIS who knew one of our coauthors (Adam Isen) and who had access to the excel spreadsheets with the data. The main barriers to access to this data (and similar data were): (1) Short institutional memory; (2) Ignorance (within the organization itself) of who has authority to release data; (3) Lack of transparency about the names of personnel and their hierarchy within government organizations.

(2) IRS data about firms

The first thing to note about IRS tax data is that the best way to access it is though coauthorship with someone in the IRS. Due to point (1)(3) above, it is difficult to know who to contact to build such a coauthorship relationship, but IRS economists who have recently authored publications are the right place to start. The second thing to note about IRS data is that IRS has easy-to-access computer files for recent tax years from 2001 onwards, but the data becomes more difficult to obtain and spottier for earlier years.

Permissions to disseminate results from IRS data are not hard to obtain if one is working with an internal IRS employee on the paper, as these employees can advise you as you work on what is allowed to be released and what is not. The bigger challenge is that the demands on IRS employees' time for policy work are extensive, resulting in delays in research projects. One solution is to obtain access as a visitor to the IRS, so that one can more directly participate in the empirical work.

**Opportunities for Collective Action to Improve the Availability of High-Quality data for the Research Community:**

(1) Funding the USCIS to record all H-1B lottery participants, not merely winners

It should be possible to make a deal with USCIS for them to record the information on all H-1B lottery participants, not merely the winners. The plan would be roughly as follows:

(i) We contact the USCIS to jointly determine an estimate of how much time it would take their employees (or contractors) to enter the I-129 application information for the losers in the H-1B lotteries.
(ii) We continue conversations with USCIS to put a dollar value on that time.
(iii) We take these estimates and a letter of intent from the USCIS to funding agencies such as the Sloan Foundation.
(iv) Having obtained funding and data, we work with this data on initial projects before establishing a protocol for sharing the data more widely with the general research community.

(2) Much better information on the **stock** (rather than the flow) of visa holders at any point in time.

One of the chief impediments to understanding the impact of high skilled immigrants on the U.S. economy is that we don't actually know much about the *stock* of current and former visa holders. The existing surveys that give some information about this stock are not sufficiently large and sufficiently regular to provide complete information, and in no cases is it easy to combine such surveys with other information at the individual level.

Only an ambitious proposal can move beyond the inadequate piece-meal data sources we currently have. What we need is a new system for keeping track of every current immigrant and non-immigrant (such as H-1B) visa-holder. If this system existed, it would be relatively easy to combine micro data from this stock at one point in time to micro data from other sources at other points in time to track the flow rate of former visa holders into ever-greater assimilation into the American economy. The key is to have information on the stock itself, at some regular temporal basis (quarterly, annually, etc).

Once again, as in (1) above, the relevant agency to work with is USCIS. The USCIS itself has complete information on the flow of immigrants into visa programs. The Department of Homeland Security (DHS) should have information on whether visa holders are actually present in the United States during any specific time period. Since the USCIS is part of DHS, the USCIS is probably the contact to start with.

Appendix:

(1) Layout of FY 2006 FY 2007 lottery data

Each row is a separate I-129 application for an individual person's H-1B visa.

The columns are as follows:

| | |
|---|---|
| LOTTERY | (whether the application won the lottery or not) |
| CATEGORY | (regular or masters desgree) |
| FY | (which fiscal year) |
| TAX_NUMBER | (employer's tax id) |
| FIRM_NAME | (name of employing firm) |
| STREET | (street address of employing firm) |
| CITY | (city of employing firm) |
| STATE | (state of employing firm) |
| ZIP | (zip code of employing firm) |

Please submit your memo by Wednesday, March 25th (ideally by 9am) using the NBER portal [http://conference.nber.org/sched/HSIs20]