# Replication steps for cross country comparisons of within-industry wage dispersion

Brad Hackinen, November 2014

This document describes how to generate the US/Germany rank comparison and the table of wage and income rank coefficients for within-industry dispersion measures.

The wage and income data comes from the Luxemburg income Study (LIS): <http://www.lisdatacenter.org/>. For privacy reasons, LIS data can only be accessed remotely via the web. Essentially, you send statistical commands to LIS, and they send you back results. The easiest way to do this is using a custom built interface provided by the institution called “LISSY”.

The following steps will replicate the analysis in the paper (unless the LIS system has changed since writing these instructions):

1. **Obtain a user account for LISSY.** Go to <http://www.lisdatacenter.org/> and follow the “register” link. Obtaining an account is free for most users, but it might take a few days for the institution to review your request.
2. **Login to LISSY.** Go to <http://www.lisdatacenter.org/> and click the Login link. This downloads a java application to your computer. Run the java application, and log in using your LISSY account information.
   * If you have problems at this stage it may be because java is out of date on your computer. Try updating java and restarting LISSY. Once problems occur it is often necessary to clear Java’s temporary internet files before running LISSY again. See <https://www.java.com/en/download/help/plugin_cache.xml>.

Once logged in, select “LIS” as the project, and “Stata” as the statistical package in the drop-down menus at the upper left. Now arbitrary Stata code can be executed with LIS data by copying and pasting Stata code to the main LISSY text field and clicking “submit”. ***NOTE: Stata commands which have the potential to return individual data are banned. See*** <http://www.lisdatacenter.org/data-access/lissy/best-practices/>

1. **Collect wage dispersion data.** Unfortunately, this is a bit complicated. Some points to understand:
   * The LISSY interface makes it difficult to fully automate the data collection process in a streamlined way. LISSY can only return text, so output must be manually copied and pasted into another file for analysis. In this case, Excel was used as an intermediate format for its convenient text import capabilities (Stata can probably do something similar). Then the excel tables were copied into Stata tables and saves as .dta files.
   * The paper presents wage dispersion measures for both income and wage, for countries with data for individual country-years, and pooled country-years. These are calculated with four different scripts:
     1. Pooled income measures:

**LISSY\_industryIncomeDispersionRank\_Pooled\_01.do**

* + 1. Single country-year income measures: **LISSY\_industryIncomeDispersionRank\_01.do**
    2. Pooled wage measures: **LISSY\_industryWageDispersionRank\_WithSector\_Pooled\_01.do**
    3. Single country-year wage measures: **LISSY\_industryWageDispersionRank\_WithSector\_02.do**
  + All four scripts share a similar structure:
    1. First, the scripts loop through a list of country-year datasets in the LIS database. For each database, they load a set of individual variables and household variables (accessed from separate LIS files).
    2. The industry codes are converted to a common 2-digit ISIC3-based format. Conversion is country specific – some countries require no change, others require extracting the middle two digits out of a 4-digit code, while a small number require applying a crosswalk. The required crosswalks are embedded directly in the script as a series of conditional replacement commands.
    3. The script calculates a set of dispersion measures
    4. The script regresses wages on a set of simple explanatory variables, and then recalculates the same set of dispersion measures for the residual. All residual-based measures have the same variable name as their raw counterpart with the addition of a prefixed “r\_”.
    5. The script calculates and prints a table of ranks according to a selected output variable. This output variable is set as a local variable near the beginning of the script.

Data collection steps:

1. Open **LISSY\_industryIncomeDispersionRank\_Pooled\_01.do**
2. Find the line where the local variable outputVar is assigned (near the beginning of the script), and set the value to “sdwage” (or r\_sdwage, as an alternative).
3. Copy and paste the script text into the LISSY interface (make sure you are on the “job session” tab). Run the script and then click the “today jobs” tab. Wait for the job to be returned. When the job is posted in the list, select it and view the “listing” tab on the right-hand side of the page. This is the script’s output.
4. Copy and paste the table data into a blank Excel notebook. Use the text to Columns tool to create a proper table (some manual cleaning up is usually required to get rid of junk rows and columns).
5. Open **LISSY\_industryIncomeDispersionRank\_01.do** and repeat steps ii-iii.
6. Paste the new data into the excel table and clean it up so that the ISIClab column is shared and each column gives the ranks for either a single country-year or set of pooled country-years.
7. Open a new Stata window and paste the data into the Data Editor. Save the result as **sdincomeRanks\_ISIClab.dta**.
8. Repeat steps i-viii, with outputVar set to “r\_sdwage”. Save the result as **r\_sdincomeRanks\_ISIClab.dta**.
   * 1. Repeat steps i-viii with L**ISSY\_industryWageDispersionRank\_WithSector\_Pooled\_01.do** and

**LISSY\_industryWageDispersionRank\_WithSector\_02.do**, saving the results as

**sdwageRanks\_ISIClab\_PrivateSector.dta** and **r\_sdwageRanks\_ISIClab\_PrivateSector.dta**.

1. **Create US/Germany comparison plots.** These four plots are created by one script. Open **rankPlots01.do** in Stata. Change the initial cd command to point to the folder where you have saved the rank data files. Run the script to generate the four plots.
2. **Create Industry S.D. rank regression slope coefficients.** The coefficients for this table are created with a single script. Open **RankCoefficientTables.do** in Stata. Change the initial cd command to point to the folder where you have saved the rank data files. Run the script to generate the coefficient values as text output. Some is fiddling required for clean LaTeX presentation.

## Expansion and modification

If the reader wishes to expand or modify the analysis described above, here are a few things to keep in mind:

* LIS datasets are stored at a country-year level. Each dataset uses the same variable naming convention, but many variables are missing from individual datasets. So adding countries is not trivial. The author attempted included every country with the necessary variables at the time when this research was first done (spring 2014). **AvailabilityMatrix.xlsx** is an extended version of the standard documentation from LIS. Scroll down to the very bottom to see binary codes on which countries have suitable variables for which analysis.
* As more recent datasets come online it may be possible to add countries or years to the dataset. If country-year datasets contain the appropriate variables, they can simply be added by name (in ccyy format) to the local variables listing included datasets at the beginning of the scripts. Datasets missing necessary variables will be automatically ignored by the script when they throw an error.
* The output variable does not have be set to sdwage or r\_sdwage. The scripts can also calculate ranks based on the following alternative measures: cvwage (the c.v. of wage), wagepctile (the 5-95 percentile range), indLabourShare (industry labour share), and their corresponding “r\_” counterparts.

## LaTeX for Description, Figure, and Table

\subsection{Cross-country patterns in industry wage and income dispersion}

We establish that the ranking of within-industry wage dispersion follows a consistent pattern across countries by using the Luxembourg Income Study Database (LIS). The LIS provides a set of cross-sectional datasets describing household and individual income and other characteristics for a large number of countries and years. These datasets have been harmonized (to the extent possible) to make variables directly comparable across countries and years.

Unfortunately, while the variables provided in the LIS are comparable across datasets, many variables are only available for a limited number of datasets. In particular, there is insufficient wage-by-industry data for Australia, Canada, Japan, Korea, Norway, Sweden, and Switzerland. However, good data exists over multiple years for the US, Germany, and Ireland. A handful of EU countries have a single year with sufficient data. The UK and France\footnote{Income data for France is not perfectly comparable to the other countries in the sample because it is reported after certain deductions are made. However, we feel that even imperfect data can be informative so we include it nonetheless.} lack wage data, but have data on total labour income. Therefore, we present two comparisons of within-industry wage dispersion: one based on total labour income, and the other based on hourly wage.

For income statistics, our sample includes all individuals between the ages of 16 and 65 with non-zero wages. Individuals are weighted by the population weight provided by the LIS. For wage statistics, we are able to restrict the sample to individuals who are employed in private industry, and we weigh individuals by their average (weekly) hours worked as well as their population weight. We are unable to consistently identify self-employed individuals, so they remain in both samples.

Our 30 industry classification system is based on 2-digit ISIC3 industry codes. Many datasets in the LIS include 2-digit ISIC industry codes or a compatible classification as a part of their labour statistics, making conversion straightforward. The US and France use unique classification systems. For these countries we constructed our own crosswalks based on the documented descriptions of the industry codes.

We rank industries according to the standard deviation of log labour income and log wage (from highest to lowest), both with and without controls. This results in four different sets of ranks overall. We use the same set of controls for both labour and wage ranks. Namely: industry (by our 30 industry classification), education (a three category classification), age (with squared and cubed terms), sex, and region (state or province, depending on the country).

We calculate the four rankings for the US and for a set of EU countries, according to data availability. Estimating standard deviations for each industry requires a large number of individuals in every industry. The LIS datasets sometimes have a very small number of individuals in particular industries. This leads to less reliable estimates of ranks. However, ranks are unlikely to shift substantially for each country over a short span of time. Therefore we also calculate pooled ranks for a few countries where this is possible. More specifically, we pool individuals in every viable dataset between 1999 and 2014 for the given country and add year dummies to the set of controls.

Figure \ref{fig:4} shows the relationship between ranks for US and Germany based on pooled samples. There is a strong correlation between ranks, regardless of which of the four statistics we use. Rankings based on income dispersion are almost identical between the two countries. The correlation between wage dispersion ranks is comparatively weaker but remains quite strong and significant. This is partly because the sample sizes are quite a bit smaller, which adds noise to the rank measures. In both cases, adding controls reduces the strength of correlation, indicating that common demographic patterns explain a small part of the correspondence between rankings. However, the fact that even the ranks calculated from wages with controls show a significant positive correlation provides robust evidence that there is some unobserved feature of industry structure which characterizes each industry and is common across countries, influencing wage dispersion.

The widespread nature of this pattern is documented in table \ref{tab:tab-wage-dispersion}. In this Table we replicate the exercise illustrated in figure \ref{fig:4} for a larger set of countries. Ranks are computed for each dataset (or pooled dataset), and then each rank is regressed individually on the corresponding ranks from the pooled US sample.\footnote{We drop the ``residual'' industry category, as it is sometimes an outlier and is not particularly informative.} Table \ref{tab:tab-wage-dispersion} reports the resulting slope coefficients and standard deviations for OLS regressions with robust standard errors.

The results in Table \ref{tab:tab-wage-dispersion} show that the industry ranks calculated for all countries in the sample are highly correlated. As in the case of Germany, the correlation for ranks of labor income dispersion is stronger than for wage dispersion. Yet, in the whole sample, controlling for demographic characteristics has a small effect, and does not always decrease the strength of the relationship. This suggests that the main source of these correlations is some other unobserved industry specific effect. Again, this evidence supports the argument that the international pattern of within-industry wage dispersion must be due to some aspect of industry structure.

Our hypothesis is that this pattern reflects different degrees of skill complementarity in the production process. While countries may be using slightly different production technologies to produce similar goods, these technology may all share some features which broadly shape the organizational structure and types of labour employed: low skill-complementarity technologies allow firms to hire a mix of high and low skill workers, while high skill-complementarity technologies encourage a more homogenous workforce, as shown in Bombardini, Gallipoli and Pupato (2012). This is consistent with the observed covariation in the international patterns of wage dispersion by industry.

\begin{figure}[DE-US-ranks]

\centering

\caption{Comparison of US and German industry ranks according to different dispersion statistics. "Residual" refers to the remaining unexplained variation after controlling for industry, education, age, sex, region (state or province), and year. \label{fig:4}}

\subfigure[Log labour income dispersion ranks]{\includegraphics[scale=.56]{sdincomeplot.eps}}

\subfigure[Log labour income residual dispersion ranks]{\includegraphics[scale=.56]{r\_sdincomeplot.eps}} \\

\subfigure[Log wage dispersion ranks]{\includegraphics[scale=.56]{sdwageplot.eps}}

\subfigure[Log wage residual dispersion ranks]{\includegraphics[scale=.56]{r\_sdwageplot.eps}} \\

\end{figure}

\begin{table}[tab-wage-dispersion]

\centering

\caption{This table reports the correspondence of within-industry income and wage dispersion patterns across countries. Values are calculated in two steps: First we calculate the standard deviation of each (logged) variable for each industry in a given country, and rank industries in that country from highest to lowest dispersion. Then the ranks of each country and dispersion statistic are regressed independently on the corresponding ranks calculated for the US using OLS with robust standard errors. The table reports the slope coefficient associated with each regression. Standard errors are in brackets. Missing values occur where the LIS has insufficient data on wage. }

\label{tab:tab-wage-dispersion}%

%\begin{tabularx}{0.9\textwidth}{c \*{2}{Y}}

\begin{small}

\begin{tabular}{c c c c c}

&\multicolumn{4}{c}{Industry S.D. rank regression slope coefficients } \\

&\multicolumn{4}{c}{\footnotesize{Dependant variable: industry S.D. ranks, pooled US 2004,2007,2010} } \\ [-1.5pt]

\cline{2-5}

& log labour & log labour & log wage & log wage \\

& income & income residual & & residual \\

& (1) & (2) & (3) & (4) \\

\cline{2-5}

\underline{\footnotesize{\textbf{Pooled country-years}}} & & & & \\

Germany 2004,2007,2010 & 0.92 & 0.88 & 0.45 & 0.45 \\ [-1.5pt]

& (0.00) & (0.00) & (0.03) & (0.03) \\ [-1.5pt]

Ireland 2000,2004,2007,2010 & 0.90 & 0.93 & 0.29 & 0.58 \\ [-1.5pt]

& (0.00) & (0.00) & (0.03) & (0.02) \\ [-1.5pt]

UK 1999,2004,2007,2010 & 0.90 & 0.89 & & \\ [-1.5pt]

& (0.00) & (0.01) & & \\ [-1.5pt]

\underline{\footnotesize{\textbf{Individual country-years}}} & & & & \\

US 2010 & 1.00 & 0.99 & 0.86 & 0.93 \\ [-1.5pt]

& (0.00) & (0.00) & (0.01) & (0.00) \\ [-1.5pt]

Germany 2010 & 0.93 & 0.89 & 0.47 & 0.61 \\ [-1.5pt]

& (0.00) & (0.00) & (0.02) & (0.02) \\ [-1.5pt]

Ireland 2010 & 0.91 & 0.88 & 0.48 & 0.59 \\ [-1.5pt]

& (0.01) & (0.01) & (0.03) & (0.02) \\ [-1.5pt]

UK 2010 & 0.56 & 0.52 & & \\ [-1.5pt]

& (0.02) & (0.02) & & \\ [-1.5pt]

France 2005 & 0.61 & 0.67 & & \\ [-1.5pt]

& (0.03) & (0.02) & & \\ [-1.5pt]

Austria 2004 & 0.79 & 0.76 & & \\ [-1.5pt]

& (0.01) & (0.01) & & \\ [-1.5pt]

Belgium 2000 & 0.83 & 0.82 & 0.42 & 0.44 \\ [-1.5pt]

& (0.01) & (0.01) & (0.02) & (0.04) \\ [-1.5pt]

Spain 2000 & 0.81 & 0.80 & 0.26 & 0.35 \\ [-1.5pt]

& (0.01) & (0.01) & (0.03) & (0.04) \\ [-1.5pt]

Finland 2007 & 0.88 & 0.87 & 0.47 & 0.56 \\ [-1.5pt]

& (0.01) & (0.01) & (0.03) & (0.02) \\ [-1.5pt]

Greece 2004 & 0.79 & 0.82 & & \\ [-1.5pt]

& (0.01) & (0.01) & & \\ [-1.5pt]

\hline

\end{tabular}%

\end{small}

\end{table}