

Stockholm, December 17, 2012

Dear Editors,

With this letter, we would like to provide a short background concerning our interaction with the authors and additional information that we discovered during the process of replicating the Dahlberg *et al.* study.

First, a little background; Nekby was part of the examination committee when Helene Lundqvist defended her thesis in October 2011. At this point, Nekby noticed and commented on a number of inconsistencies in the study especially concerning the description of and the use of the refugee placement policy as an exogenous source of variation. We both attended a seminar in November 2011 at Stockholm University where Matz Dahlberg presented the paper. At this point, we were also concerned with the few observations from the survey data used in the Dahlberg *et al.* analysis. We argued, that they should use all available data since their identifying variation is at the group level (the municipal level in this case). Our concerns about these issues, and what it may imply for their results, lead us to take a closer look at their study.

The following is a short chronological outline of our replication of Dahlberg *et al.* (see also a summary of key findings provided below):

- We requested access to the data from the authors (e-mail request sent on November 22, 2011). The authors responded that they were not allowed to distribute data due to requirements stipulated in the contracts signed with the Swedish National Data Service (SND) (e-mail reply December 3, 2011). However, since SND has no property rights over these data we contacted the primary researcher for the Swedish National Election who granted us permission to make this data publicly available. Thus, their claim that the data are proprietary is false.
- Since we could not get the data from the authors, we decided to collect the data ourselves. Once we had collected the data and began the re-analysis, we were unable to replicate any of the results presented in the Dahlberg *et al.* study based on the information available in the article.
- We wrote to the authors again explaining the problems that we encountered in attempting to replicate their study and they eventually sent us their do-files (on April 27, 2012) but still not their data.
- When we analyzed their do-files we discovered (1) that there were a large number of unreported and unwarranted sample restrictions (the rotating panel data sample is reduced from 2702 to 1917 observations, leading to an attrition rate of 66% relatively to the original sample of 5571) and (2) that the authors wrongly claim that their instrument reflects the number of refugees determined by the placement program. In actual fact the instrument Dahlberg *et al.* use is based on state grants paid to municipalities to cover expenses for all refugees registered in the municipality. These two measures are not the same as the latter includes a large number of refugees in municipalities that were *not* placed there as part of the placement policy.
- In collecting data, we discovered that the Swedish Immigration Board (SIV) published two series concerning refugees; (1) the contracted levels negotiated between SIV and municipalities (due to the placement policy) and (2) the series used by Dahlberg *et al.* based on state grants for actual refugee settlement. In their paper, however, Dahlberg *et al.* state that their instrument is the number of “refugees placed within the placement program”. This is obviously not the case since state grants also cover tied-movers,

refugees who have migrated internally within Sweden and asylum seekers, none of which are placed in the municipality via the placement program.

- We sent the authors our replication study on June 28, 2012. Thereafter, the authors contacted us on August 24 asking for our data and do-files, which we promptly provided (August 26). At this point, we again asked for their data, which they finally sent (excluding information from the Swedish Election Studies) on September 3.
- With access to their data, we discovered that their data do not come from the cited sources but rather, at least in part, from colleagues and other unreported sources. This is not stated in the original paper but acknowledged in their response to our replication (footnote 4 in their comment). In comparison to the original data (which we collected from the archives of SIV and Statistics Sweden) there are a large number of inconsistencies in the Dahlberg *et al* data, which we describe in greater detail below.
- After several requests (we were initially informed that we would get a response in late August), we finally received their written response on November 26.

Inconsistencies with the Dahlberg *et al* data:

As mentioned above, Dahlberg *et al.* claim that data cannot be posted on JPE homepage. Instead, the authors have posted the data at the Swedish National Data Service (SND) (as *SND 0906*). In order to gain access to this data, a formal request at SND is required, subject to the approval of the authors, thus essentially giving them veto rights on any attempts at replication. We would like to stress here, that the material posted by Dahlberg *et al.* only includes the data necessary to re-produce their reported results (i.e., after sample restrictions and data transformations etc. have been made) and not the full data material available. This implies that it is not possible, without going back to the original survey data and with additional extensive data collection efforts, to gain an understanding of the robustness of the results presented in their study.

Our comparison of the original data, collected by us from the archives of the Swedish Integration Board and Statistics Sweden, with the data used by Dahlberg *et al* indicates the following inconsistencies with the Dahlberg *et al* data:

Instrument (data source: the Swedish Integration Board)

- There are 24 inconsistencies in their data. Particularly noteworthy is that there are observations based on the *contracted* number of refugees rather than actual refugee settlement as measured by state grants.

Housing vacancies (data source: Statistics Sweden)

- Data from the municipalities Haparanda, Pajala, Valdemarsvik, Borgholm, Lomma, Grästorp, Gnesta, Trosa are coded as missing observations when the correct coding is zero.
- Data for the municipalities of Mullsjö and Habo are coded as missing observations in 1994 when information for these municipalities and year is available.

Welfare spending (data source: Statistics Sweden)

- Data from the municipalities Alingsås, Burlöv, Gävle, Hudiksvall, Hultsfred, Härnösand, Härryda, Mariestad, Sotenäs, Stenungsund, Södertälje, Trosa, Täby and Örnköldsvik are coded as zero when they should be coded as missing observations.

Immigrants (data source: Statistics Sweden)

- We discovered that there are systematic inconsistencies in the data on immigrants defined by citizenship. In particular, the number of individuals with unknown

citizenship differs for the data used by Dahlberg *et al.* The data provided by Statistics Sweden has more immigrants classified with unknown citizenship. We have no explanation for why the Dahlberg *et al.* data do not match the citizenship data provided by Statistics Sweden.

- We also discovered that Dahlberg *et al.* lack data on immigrants (according to citizenship) for the year 1985 and replace the missing data by taking an average of the information available in 1984 and 1986. This is not discussed in their paper. Again, it is important to stress that this “missing” information is available at Statistics Sweden.

Summary of Key Findings from our Re-analysis:

In our comment we argue that the results in Dahlberg *et al.* (2012) is based on (i) an endogenous instrument and (ii) severe sample attrition bias

- Dahlberg *et al.* define their instrument—the Swedish refugee placement program—as actual refugee settlements or migration, which is clearly an endogenous outcome. Instead, we argue that the Swedish refugee placement policy should be based on the written contracts between the municipalities and the Swedish Immigration Board. With our definition of the instrument, there is *no* relationship between ethnic diversity and preferences for redistribution regardless of estimation sample used.
- It is a well-know fact that when the regressor of interest varies at a more aggregate level such as state or municipality, individual panel data is not required for identification. Despite this fact, Dahlberg *et al.* insist on analyzing data only from the rotating panel (individual-level data), which has a much large attrition rate/non-response rate than the repeated cross-section sample. Furthermore, they make additional sample restrictions on the rotating panel, further increasing the attrition rate to 66%. We show that when estimation is based on samples with considerably less sample attrition, the estimated coefficients are reduced considerably and are no longer significantly different from zero in the sample with the smallest attrition rate (see Tables 5 and 6). We also show when the regressions are re-weighted with population weights to reflect the population regression of interest (see Tables 7 and 8) there is no effect even in the sample with the largest attrition rate, i.e., the sample used by Dahlberg *et al.*

We also note that Dahlberg *et al.* fail to provide key pieces of information from the published paper thus making it impossible to evaluate the robustness of their empirical results without truly expert knowledge of the Swedish setting.

- They provide a very biased description of the workings of the refugee settlement policy (see section 2 of our comment). Specifically, they only cite papers (Edin *et al.*, 2003 and Bengtsson, 2004) that support their argument while there are numerous other papers that tell a different story. Moreover, they fail to test to what degree housing vacancies are a “key determinant of refugee settlement”, which they argue makes refugee placement conditionally exogenous. The fact that housing vacancies are not correlated with the placement policy casts doubt on their descriptions of the placement policy (see Table 3 in our comment).
- Dahlberg *et al.* write that that their instrumental variable “**refers to refugees placed within the placement program.**” However, there is *no* information on how the instrumental variable is defined in their paper. We discovered that their instrumental variable is based on actual refugee settlement which is clearly not the same as “refugees placed within the placement program” since actual refugee settlements also

include a large number of refugees not part of the placement policy such as tied-movers, resettled refugees, asylum seekers etc. The information provided in the statistical publications of the Swedish Immigration Board makes clear that actual refugee settlement is based on State remuneration to municipalities for refugee settlement, regulated according to Decree 1984:683 and 1990:927. Moreover, the number of contracted refugees is typically published in the same table as actual refugee settlement in the statistical publication of the Swedish Immigration Board. Thus, there is little room for any misconceptions of what data is available and what the two series measure.

- No information about the statistics of the rotating sample is provided, i.e., the size of the rotating panel, non-response rates, attrition rates etc. The only information given is “This study uses information from waves 1982, 1985, 1988, 1991, and 1994, when roughly 3,700 individuals were surveyed in each wave.” In our comment, we provide these statistics in Table 4 showing that the full sample size in the rotating panel is 5,571 observations but, due to attrition, only 2,703 observations are available. We also show that the size of the available repeated cross-sectional is 9,620 observations with an attrition rate of only 33%.
- There is **no** discussion of the additional (and unnecessary) sample restrictions made on the rotating panel data, which further reduces the sample size by 581 observations. This is however acknowledged in their comment to our replication. In their comment, the authors continue to base their analysis on a smaller number of observations than available since they continue to exclude those individuals that have moved between two surveys (see next point).
- There is no discussion in the original paper about the theoretical reason for excluding the 205 individuals that change municipality of residence between survey periods (see p 55). In fact, excluding observations will lead to sample selection bias. This problem can, however, be eliminated by defining comparison groups based on prior (to survey response) place of residence, as prior residence cannot be affected by the treatment (as discussed by Angrist and Pischke, 2008).
- Dahlberg *et al* define immigrants with citizenship rather than country of birth. We show that when country of birth is used the IV estimates decrease by at least 40%, which raises concerns about the exclusion restriction (see Table 9 in our paper).

Finally, we would also like to highlight inconsistencies in their response to our replication, Dahlberg et al. (2012b):

- We note that they still misleadingly refer to their instrument variable as “the number of placed refugees” despite the fact that their variable includes a large share of immigrants not part of the placement program as noted above.
- We also note that the authors continue to use exactly the same data in their response to our replication despite the numerous inconsistencies in these data. They could easily have discovered these inconsistencies since we have provided both our data and do-files. However, in footnote 10 of their response, they acknowledge that the two data sets differ stating that they “believe this is due to a few data typing errors resulting in missing values in some of our variables.”
- It is noteworthy that they come to a completely different conclusion when they investigate the relationship between the written agreements and actual settlements levels since they state that the relationship “is somewhat dependent on the time period and the type of variation used to correlate the two, but is in most specifications highly statistically significant and close to one.” However, the results from Table 1 in our

paper clearly illustrate that the relationship between actual refugee settlement and contracts are highly non-robust: depending on time period, specification of the population regression (weighting scheme) and re-definition of the policy measure (normalized with population shares).

- In their response, they argue that contracted number of refugees is a weaker instrument than actual number of refugees and therefore more biased. This argument is wrong since in a just-identified model, the estimate is approximately (median) unbiased (unless the first-stage is really weak) and therefore one cannot assess which of the two IV estimates are more or less biased. In addition, the reduced form estimate is unrelated to the strength of the first-stage estimates. Thus, if the reduced form effect is not statistically different from zero, then there does not exist a causal relationship between immigrants and preferences for redistribution.
- In their response, they also argue that the repeated cross-section data cannot be used since cell sizes are too small, with an average of about 9 observations. This argument is clearly wrong since in their rotating panel data specification the number of observations per cell is only 3. In other words, they fail to understand that their panel data specification is also a grouped-data regression since their regressor only varies at the municipality level.
- In their response, the authors still make sample restrictions in the rotating panel. They exclude those individuals that have moved between two surveys, which mean that they have only 2446 observations instead of 2702.
- They do not include any results from the repeated cross-section, which has 9,620 observations.
- Concerning the use of citizenship as a measure for ethnicity rather than country of birth, the authors write “at the time of writing the paper, we only had access to citizenship, so the alternative definition was not an option.” We note that these data are available at Statistics of Sweden both then and now.

Sincerely,

Lena Nekby and Per Petterson Lidbom