**Data Sources for "The Twin Instrument: Fertility and Human Capital Investment"**

This document describes the sources of data used to generate the results in "The Twin Instrument: Fertility and Human Capital Investment" by Sonia Bhalotra and Damian Clarke in the Journal of the European Economic Association. Additional details on replication materials should be found in the README.txt file, which describes full replication procedures for figures and tables and is bundled along with these replication materials.

Results are based on three sources of data, (1) The Demographic and Health Surveys (2) The National Health Interview Surveys and (3) the National Longitudinal Survey of Youth 1979, children and young adults panel. Below we provide a description of the process to generate data files used in the paper which are provided with these replication materials.

**The Demographic and Health Surveys (DHS)**
We downloaded all publicly available Demographic and Health Surveys from 1990 onwards, up until 2012. We match all births occurring to mothers from the DHS Birth Recode (BR) with their later educational attainment from the DHS Person Recode (PR) files. These records are matching using the DHS ID (from variables v001 and v002), as well as a survey country and survey year unique ID. The Birth Recode provides information on each birth occurring to each mother including their year of birth, sex, and twin status. Additionally, this provides mother's characteristics such as her height, weight and education. The file `DHS_twins_JEEA.dta` consists of all children living in the same household as their mother. Main results in replication files are drawn from the sample aged between 6 and 18 (ie those for whom education is recorded and who are 18 years of age or younger).

**The National Health Interview Surveys (NHIS)**
We downloaded each dataset for each National Health Interview Survey between 2004 and 2014 inclusive, publicly available from the United States' Center for Disease Control and Prevention (CDC). In order to generate the analysis file used in this paper (`NHIS_twins_JEEA.dta`) we used data from the family file (`familyxx.dta`), the household file (`househld.dta`), the person file (`personx.dta`) and the sample adult file (`samadult.dta`). These are all released as standard by the NHIS in each year under study, and were merged using the NHIS ID in each year, generated from the variables `hhx` (household identifier), `fmx`) (family identifier) and `fpx` (person identifier). The file provided with replication materials contains all children matched with their mothers in these files.

**The National Longitudinal Survey of Youth 1979 (NLSY79) Children and Young Adults Panel**
Finally, we extract information for each respondent of the NLSY79 Children and Young Adult Panel, freely available from the NLS Web-Investigator. The full Tagset of variables extracted along with their NLSY79 labels is provided as the plain text file NLSYchild.CHILDYA in the data subfolder of these replication materials. This can be used to extract original data directly from the NLS Web-Investigator. This original data is then reshaped to form a file with one line per child and mother with information on each investment variable described in the paper, along with the child's birth year, twin status, and school outcomes. The resulting dataset for use in all analysis is named `NLSY79-kids.dta`.