Data construction explanation for the paper: "Neighbourhood Turnover and Teenage Attainment", by Stephen Gibbons, Olmo Silva and Felix Weinhardt, *Journal of the European Economic Association.*

The datasets used are tabulated below:

Data	Description	Provider and access	Used in
National Pupil Database 1998-2008 including Pupil Level Annual School Census (PLASC) 2002-2008	Demographics, postcode-of-residence and information on school attended in every year for children in England. Key Stage 1 test scores 1998-2001 Key Stage 2 test scores 2002-2005 Key Stage 3 test scores 2005-2008	Department for Education and Skills: NPD.Requests@education.gsi.gov.uk Controlled access. Subject to approval, these data are available for use by any organisation or person who, for the purpose of promoting the education or well-being of children in England. Requires secure storage arrangements	All tables except Table 8
Edubase 2002-2008	School institutional information, such as religious affiliation, postcode of location (for distance calculations in GIS), type of institution (e.g. community, voluntary aided, voluntary controlled, foundation school)	Department for Education and Skills: http://www.education.gov.uk/edubase/home.xhtml Historical data available on request from Department for Education and Skills	All tables
Longitudinal Study of Young People in England (LSYPE)	Panel study of one cohort of children aged 13-14 in 2004. Includes variables to construct student-level behavioural outcomes. Can be linked to NPD by agreement with Department for Education and Skills	UK Data Archive https://discover.ukdataservice.ac.uk/series/?sn=2000030 Department for Education and Skills	Table 8
Land Registry Price Paid data	Micro data on residential house sales transactions from 1995 to present	Publicly available https://www.gov.uk/government/collections/price-paid-data	Table 3, Table 6
UK Population Census 2001	Aggregated population census data. Data at Ouput Area level.	Publicly available (to UK-based researchers) from the Office for National Statistics and through various sources (http://nomisweb.co.uk, http://casweb.mimas.ac.uk/, https://discover.ukdataservice.ac.uk/catalogue/?sn=5835&type=Data%20catalogue)	Table 6
General Practice patient	Data on patients registered at National Health Service	Health and Social Care Information Centre	Table 7

register	GP practices in England for 2002-2008. Contains data on total capitation, registration and deregistration of patients.	Bespoke purchase. Contact details available at: www.hscic.gov.uk	

In the following, we describe how we combined these datasets to undertake our main analysis. The Do-Files for the replication of the results of the paper are also posted online.

From PLASC, the following information was kept: pupil identifiers, school identifiers, Local Education Authority (LEA) identifiers, ethnicity, special education needs (SEN) eligibility for free school meals (FSME), gender, academic years (e.g. 2001/2002), and school year/grade (e.g. 6^{th}, 7^{th} grade).

Several years of PLASC were used to track every pupils' school, postcode of residence and SEN and FSME status for school grades 6^{th} to 9^{th}.

Next, the following data from NPD was used: pupil identifiers, LEA identifiers, KS1 English and Math levels, KS2 total point scores and levels in English, Maths and Science, and KS3 total point scores and levels in English, Maths and Science (to construct KS2 and KS3 cohort-specific percentiles in the national distribution). Note that, using levels of achievements, we recode some originally missing total point score entries in order to assign them to a percentile. In particular, at both KS2 and KS3 and for the three subjects, students with the following test entries are assigned to a zero: absent; below the test level (not admitted to the test); dis-applied from the test (not admitted to the test); not awarded a valid scores (very low performers); working towards the test (very low performance); malpractice (misbehaviour during the test). We then go on to construct percentiles in the cohort-specific national distribution of KS2 and KS3 total point scores. Note also that KS3 total point scores were adjusted for 'tiering' before creating percentiles. More details on the exact procedure to account for 'tiering' can be obtained from the authors.

The following information was extracted from the Edubase years 2002/2008: school and LEA identifiers, school denomination (e.g. Catholic, Church of England, None), school admission policies (e.g. Comprehensive, Grammar), school specialism, school capacity, school phase (e.g. Primary, Middle-Secondary, Secondary) and school type (community, voluntary aided, voluntary controlled, foundation). This information is required to drop students attending "selective" schools, as described in Section 2 of the paper.

Information on pupils from NPD and PLASC was merged using pupil identifiers. Subsequently, school information was added to students' background information and test scores using data from Edubase for the relevant years using school identifiers.

Using these data, a panel was constructed including pupil-level information for four cohorts of students for every school-year between grade 6 (end of primary school) and grade 9 (middle of secondary school). In this initial set-up, information for pupils' test scores in different subjects (at KS1, KS2 and KS3) is entered as different columns. Time-fixed background information is also entered column-by-column (one column per characteristics) and time-changing individual and school characteristics (e.g. postcode of residence and total school roll) are entered as different columns. Note

that we keep pupils for whom we can reconstruct full information on schools attended, test scores and background characteristics for all the years in our analysis. This means we keep students that appear in every wave of PLASC and NPD, to whom we can match school information and for whom pupil and school information is non-missing. As a result, our measures of neighbourhood mobility (see below) do not result from students dropping out of the data, or appearing only in certain years.

(Note that for the analysis using the LSYPE outcome variables in Table 8, we had to extract and create mobility measures for a "fifth" cohort that we do not include in the main analysis due to the lack of KS1 information for years prior to 1998.)

This data is the starting point for the creation of our working dataset. The following main additional steps were undertaken.

First, we compute neighbourhood-specific mobility measures. To do this, we match the postcode information to the 2001 Census Output Area definitions to delimit neighbourhoods and exploit the three transitions from academic years 6 to 7, 7 to 8 and 8 to 9 in order to construct cohort-specific counts of students leaving and moving into each neighbourhood. These are then averaged over the three years of each cohort and age as described in Section 2 of the paper. Our main analysis focusses on neighbourhood turnover of same-age children, e.g. for the transition year 6 to year 7 we count moving pupils who belong to the same academic year/cohort. However, for the additional analysis in Table 7, we also construct mobility measures based on mobile students or mobile adults belonging to other age groups, by student and mover characteristics and by school attended.

Second, we use the panel obtained from PLASC to construct variables that capture the compositional changes of average neighbourhood characteristics that are induced by student mobility. These are summarised in Table 1 of the paper and included as controls in our regressions.

Finally, we merge the three datasets (PLASC extract, OA-level mobility stats, Pupil-level neighbourhood composition) together and add information at the OA-level on house prices from the Land Registry and Output Area characteristics from the UK 2001 Census of Population (used in Table 3 and Table 6), as well as the mobility rates of adults from the NHS data (used in Table 7).

In terms of sample selection, we keep students in non-selective secondary schools (see Section 2), and drop neighbourhoods with less than 5 students overall or "extreme" mobility measures of over 100 percent. The main analysis is then further restricted to students who do not change their postcode between academic years 6 and 9, whom we define as 'stayers'. Note that in Column (6) of Table 4 we extend our analysis by keeping the movers.

Using this dataset, we then run our main analysis where we regress stayers' value added on various neighbourhood mobility, characteristics and controls. This is detailed in the do-file that is posted alongside this document.

We create and merge similar information (adjusted to the relevant cohort) to the LSYPE data extract and perform our main regression analysis using the second do-file posted along with this document.

The data manipulation and econometric analysis were carried out in Stata13© Multi-Processor on a 64-bit machine with 192GB RAM.