School Segregation Data Linked to Next Steps (Sweeps 1 to 8)

User Guide (Version 1)

May 2021

Contacts

If you have any queries, please contact us at

[nadia.siddiqui@durham.ac.uk](mailto:nadia.siddiqui@durham.ac.uk)

[shao.xin@durham.ac.uk](mailto:shao.xin@durham.ac.uk)

How to cite this guide

Shao, X. and Siddiqui, N. (2021). School Segregation data Linked to Next Steps (Sweeps 1 to 8). [Data Collection]. Colchester, Essex: UK Data Service.

Acknowledgements

Special thanks go to the fantastic colleagues at the Centre for Longitudinal Studies (CLS), UCL – including, but not limited to, Aida Sanchez and Sarab Rihal.

Contents

[1. Introduction 1](#_Toc72318423)

[2. School segregation indices 1](#_Toc72318424)

[3. Data linkage to Next Steps 2](#_Toc72318425)

[4. The research dataset 3](#_Toc72318426)

[5. School Segregation Research Report 5](#_Toc72318427)

[5.1 Introduction 5](#_Toc72318428)

[5.2. Data and data linkage 6](#_Toc72318429)

[5.3. Potential research areas 9](#_Toc72318430)

[5.4. Data access 9](#_Toc72318431)

[6. References 10](#_Toc72318432)

# 

# 1. Introduction

Next Steps is a longitudinal cohort study following a nationally representative group of nearly 16,000 people born in England in 1989-90. The study began when the cohort members were 14 years old. With sweeps every year for the first seven years, it has captured incredibly rich information about their educational trajectories during adolescence and early adulthood.

Today, Next Steps continues to chart this cohort’s experiences with a wider disciplinary scope, providing unparalleled insights into the many different aspects of their lives and transitions into adulthood.

A vital source of evidence, Next Steps has had a major influence on national education policy and cast light on a wide range of important social issues, including the effects of zero hours contracts and bullying.

Next Steps is a large-scale nationally representative longitudinal study and important new data resource for addressing research questions regarding various social issues. Education administrative data from the National Pupil Database (NPD) and the Individual Learner Records (ILR), both provided by the UK Department for Education, have been linked to the Next Steps survey data and are publicly available from the UK Data Service.

The National Pupil Database (NPD) collects administrative data from all state-maintained schools in England and provides rich school-level data including the specific schools that pupils attend, their educational attainment and “a range of possible indicators of pupil disadvantage such as eligibility for free school meals and special educational needs,” which can facilitate analysis of the main characteristics of school segregation.

The school segregation dataset was created from original NPD data by researchers at the School of Education, Durham University. For detailed information on how this dataset was generated, research background information and references, please refer to the School Segregation Research (SSR) report in the Appendix to this User Guide.

# 2. School segregation indices

‘School segregation’ refers to clustering of children with similar socio-economic backgrounds in particular schools, i.e. who goes to school with whom. School segregation measures unevenness in the distribution of student characteristics among schools. For further explanation, details and references, refer to the SSR report (Appendix).

The authors of the dataset used the publicly available school-level data from the NPD and identifiedrelevant data on the composition of disadvantaged groups in the school population of full-time equivalent students. In particular, they used six common indicators of disadvantage in schooling that have been identified as potential factors in school stratification by social group:

1) the number of pupils taking free school meals (FSMt),

2) the number known to be eligible for free school meals (FSMe),

3) the number known to have a statement of special educational needs (SENs),

4) the number known to have special needs without a statement (SENn),

5) the number known to have English as a second or additional language (ESL/EAL)

6) the number classified as non-white British ethnic origin (NW)

The relevant figures for each indicator for each school in each year were converted into what has been termed the Gorard Segregation Index (GS) and the Dissimilarity Index (D) at the national level in order to probe into the main characteristics of school segregation:

* The **Gorard Segregation Index (GS)** is an index of segregation between schools that refers to “the proportion of potentially disadvantaged students in a school system who would have to exchange schools with another (non-disadvantaged) student for there to be no segregation by disadvantage between schools in that school system (or the area under consideration)” It is calculated in a similar way to the disparity ratio used in health studies.
* The **Dissimilarity Index (D):** the residual for D is the absolute value of the result of subtracting the population proportion of non-FSM pupils in each school from the population proportion of FSM pupils in each school; D itself is the sum of these residuals for all schools divided by two.

Further descriptions of these indices can be found in Section 4.

The GS and D indices were initially calculated for all the secondary schools available for each year and each of the six school indicators listed above using figures available from the School Level Annual Schools Census (SLASC) from the National Pupil Database (NPD) for state secondary schools at the national level in England from 2003 to 2006.

# 3. Data linkage to Next Steps

The school-level segregation dataset (prepared by researchers at Durham University) was linked to the school at which the Next Steps cohort member had a matched KS4 record in the NPD via the school’s unique reference number (URN). For the majority of cases, this was at age 16. The school-level dataset was then linked to the individual cohort member via the anonymised research identifier (NSID) and the actual key stage 4 school identifier was replaced with the anonymised URN currently used in [*Next Steps: Linked Education Administrative Datasets (National Pupil Database), England, 2005-2009: Secure Access*](https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7104). The resulting dataset has one row per cohort member, with 15272 out of a total of 16122 cases in the dataset having a matched KS4 record. These linked data are now available in Secure Lab data waves 1-8.

School segregation indices for four years (2003-2006) were used for data linking for the purpose of completeness of the linked files. School segregation is a consistent phenomenon and therefore the indices do not drastically change for these years.

All the data used for calculation are at the school level and the school unique reference numbers (URNs) were *only* used for linkage. This linkage was independently conducted by the Centre for Longitudinal Studies (CLS), which was given access to the data files with matching codes for schools.

Some variables were relabelled, and the final dataset was de-anonymised by reducing the number of decimal digits.

# 4. The research dataset

The linked school segregation dataset contains 47 variables. It includes the GS and D school segregation indices for each indicator for each school in each year. Specifically, the indicators for the GS and D indices are: 1) the number of pupils taking free school meals (FSMt), 2) the number of pupils known to be eligible for free school meals (FSMe), 3) the number known to have a statement of special educational needs (SENs), 4) the number known to have special needs without a statement (SENn), 5) the number of pupils classified as white British ethnic origin (NW), 6) the number of pupils classified as non-white British ethnic origin (NW), and 7) the number known to have English as an additional language (EAL).

The variables have been appropriately named and labelled to refer to the index, the indicator and the year.

Table 2 lists all the school segregation-related variables available in the dataset.

**Table 2: Segregation data variable descriptions**

| Variable Name | Variable Label and Description |
| --- | --- |
| NSID | Anonymised Research ID |
| KS4\_URNANON | School's Unique Reference Number (anonymised) |
| LEA03\_GS\_FSMe | Gorard Segregation Index (GS) for FSM eligibility 2003 |
| LEA03\_D\_FSMe | Dissimilarity Index (D) for FSM eligibility 2003 |
| LEA03\_GS\_SENs | Gorard Segregation Index (GS) for SEN with statement 2003 |
| LEA03\_D\_SENs | Dissimilarity Index (D) for SEN with statement 2003 |
| LEA03\_GS\_SENn | Gorard Segregation Index (GS) for SEN without statement 2003 |
| LEA03\_D\_SENn | Dissimilarity Index (D) for SEN without statement 2003 |
| LEA03\_GS\_W | Gorard Segregation Index (GS) for White ethnic origin 2003 |
| LEA03\_GS\_NW | Gorard Segregation Index (GS) for Non-White ethnic origin 2003 |
| LEA03\_D\_NW | Dissimilarity Index (D) for Non-White ethnic origin 2003 |
| LEA03\_GS\_EAL | Gorard Segregation Index (GS) for English as an additional language 2003 |
| LEA03\_D\_EAL | Dissimilarity Index (D) for English as an additional language 2003 |
| LEA04\_GS\_FSMe | Gorard Segregation Index (GS) for FSM eligibility 2004 |
| LEA04\_D\_FSMe | Dissimilarity Index (D) for FSM eligibility 2004 |
| LEA04\_GS\_SENs | Gorard Segregation Index (GS) for SEN with statement 2004 |
| LEA04\_D\_SENs | Dissimilarity Index (D) for SEN with statement 2004 |
| LEA04\_GS\_SENn | Gorard Segregation Index (GS) for SEN without statement 2004 |
| LEA04\_D\_SENn | Dissimilarity Index (D) for SEN without statement 2004 |
| LEA04\_GS\_W | Gorard Segregation Index (GS) for White ethnic origin 2004 |
| LEA04\_GS\_NW | Gorard Segregation Index (GS) for Non-White ethnic origin 2004 |
| LEA04\_D\_NW | Dissimilarity Index (D) for Non-White ethnic origin 2004 |
| LEA04\_GS\_EAL | Gorard Segregation Index (GS) for English as an additional language 2004 |
| LEA04\_D\_EAL | Dissimilarity Index (D) for English as an additional language 2004 |
| LEA05\_GS\_FSMe | Gorard Segregation Index (GS) for FSM eligibility 2005 |
| LEA05\_D\_FSMe | Dissimilarity Index (D) for FSM eligibility 2005 |
| LEA05\_GS\_SENs | Gorard Segregation Index (GS) for SEN with statement 2005 |
| LEA05\_D\_SENs | Dissimilarity Index (D) for SEN with statement 2005 |
| LEA05\_GS\_SENn | Gorard Segregation Index (GS) for SEN without statement 2005 |
| LEA05\_D\_SENn | Dissimilarity Index (D) for SEN without statement 2005 |
| LEA05\_GS\_W | Gorard Segregation Index (GS) for White ethnic origin 2005 |
| LEA05\_GS\_NW | Gorard Segregation Index (GS) for Non-White ethnic origin 2005 |
| LEA05\_D\_NW | Dissimilarity Index (D) for Non-White ethnic origin 2005 |
| LEA05\_GS\_EAL | Gorard Segregation Index (GS) for English as an additional language 2005 |
| LEA05\_D\_EAL | Dissimilarity Index (D) for English as an additional language 2005 |
| LEA06\_GS\_FSMe | Gorard Segregation Index (GS) for FSM eligibility 2006 |
| LEA06\_D\_FSMe | Dissimilarity Index (D) for FSM eligibility 2006 |
| LEA06\_GS\_SENs | Gorard Segregation Index (GS) for SEN with statement 2006 |
| LEA06\_D\_SENs | Dissimilarity Index (D) for SEN with statement 2006 |
| LEA06\_GS\_SENn | Gorard Segregation Index (GS) for SEN without statement 2006 |
| LEA06\_D\_SENn | Dissimilarity Index (D) for SEN without statement 2006 |
| LEA06\_GS\_SEN | Gorard Segregation Index (GS) for SEN 2006 |
| LEA06\_D\_SEN | Dissimilarity Index (D) for SEN 2006 |
| LEA06\_GS\_W | Gorard Segregation Index (GS) for White ethnic origin 2006 |
| LEA06\_GS\_NW | Gorard Segregation Index (GS) for Non-White ethnic origin 2006 |
| LEA06\_D\_NW | Dissimilarity Index (D) for Non-White ethnic origin 2006 |
| LEA06\_GS\_EAL | Gorard Segregation Index (GS) for English as an additional language 2006 |
| LEA06\_D\_EAL | Dissimilarity Index (D) for English as an additional language 2006 |

# 

# 5. School Segregation Research Report

## 5.1 Introduction

In general, data linkage refers to matching two or more sets of data which have common cases or units and can be linked with a common identification code. Linking can be an exact matching of cases or units, but depending on the quality of the information it can be a fuzzy matching, with cases or units being linked on the basis of some similarity threshold criteria. In social research, linking secondary data sets for the purpose of research is becoming increasingly popular. It is particularly valuable because it allows analyses using additional information such as high-quality administrative information (Siddiqui 2019). Especially in recent years, linking large-scale longitudinal survey data to administrative data has become a popular trend in sociology research (e.g. Sala, Burton & Knies, 2012; Siddiqui, Boliver & Gorard, 2019).

As an innovative large-scale nationally representative longitudinal (panel) study, Next Steps has become an important new data resource for addressing research questions regarding various social issues. Using Next Steps data, research has already been conducted to address research questions including on social behaviour (Demack, Platts-Fowler, Robinson, Stevens & Wilson, 2010), social mobility (Goodman & Gregg, 2010; Note, 2019), access to higher education (Anders, 2012; Siddiqui, Boliver & Gorard, 2019) and health behaviour (Hagger-Johnson et al., 2013; Williams & Hagger-Johnson, 2017). One advantage of Next Steps data is that they have been linked to administrative data from the National Pupil Database (NPD) (for detailed information see Department for Education, 2011). We go one step further by linking school segregation data created from the NPD to Next Steps (see the next section for more information on segregation data).

## 5.2. Data and data linkage

Compared to previous data linkage practice which links two or more *original* datasets, we have pioneered a new data linkage – linking longitudinal data to school segregation data created from the original NPD data to provide evidence and reference for potential research on the wider impact of school segregation. The longitudinal survey data contain in-depth details of the participants’ attitudes to their lives, home environment and long-term outcomes which are not available in any other administrative records. The participants self-report information on their school environment and experiences at school. Understanding the association between clustering by poverty at the school level with the participants’ attitudes, aspirations and life outcomes is made possible by linking school segregation indices with the schools the study participants attended. This is now possible.

### 5.2.1. What is School Segregation and why does it matter?

‘School segregation’ refers to segregation between schools – it is a kind of clustering in which children from similar socio-economic backgrounds are segregated into particular schools, i.e. who goes to school with whom (Gorard, 2018). School segregation is measurable with characteristics of disadvantage such as poverty, ethnic origin and learning difficulties. It measures the unevenness in the distribution of student characteristics among schools, particularly in terms of disadvantage (for further explanations and detail, see Gorard, Taylor & Fitz, 2003; Gorard, 2016; Gorard, 2018).

Why does the clustering of pupils matter? Between-school segregation in terms of students’ socio-economic characteristics has been well established in studies as relating to school effectiveness and school outcomes in the form of public examination scores (Gorard, 2000). The segregation of students can also be linked to a wide range of other student outcomes, including social skills, emotional and behaviour problems, higher education participation and civic engagement, etc. (for further discussion, see Gorard, S. & See, B. H., 2013; Gorard, 2014; Gorard, 2018).

### 5.2.2. How we measure school segregation – Introducing the Gorard Segregation Index (GS) and the Dissimilarity Index (D)

Gorard, Taylor and Fitz (2003) present large-scale in-depth evidence suggesting a number of the determinants of school segregation at both the local and national levels. In a recent paper, ‘The complex determinants of school intake characteristics and segregation, England 1989 to 2014,’ Gorard (2016) gives a detailed explanation of the six main common indicators of disadvantage in schooling: 1) the number of pupils taking free school meals (FSMt), 2) the number known to be eligible for free school meals (FSMe), 3) the number known to have a statement of special educational needs (SENs), 4) the number known to have special needs without a statement (SENn), 5) the number known to have English as a second or additional language (ESL/EAL), and 6) the number classified as white British ethnic origin (NW) (see Gorard, 2016 for detailed information and discussion). The National Pupil Database (NPD) provides relevant data on the composition of disadvantaged groups in the school population of full-time equivalent students, which allows analysis of the potential factors behind school stratification by social group such as FSMt, FSMe, SENn, SENs, ESL/EAL and NW.

The relevant figures for each indicator for each school in each year have been converted into what has been termed the Gorard Segregation Index (GS) and the Dissimilarity Index (D) at the national level in order to probe into the main characteristics of school segregation. The GS is an index of segregation between schools that provides “the proportion of potentially disadvantaged students in a school system who would have to exchange schools with another (non-disadvantaged) student for there to be no segregation by disadvantage between schools in that school system (or the area under consideration)” (Gorard, 2018, p. 14). It is calculated in a similar way to the disparity ratio used in health studies.

To illustrate the measurement of the GS, we use FSM as an indicator of disadvantage as an example. This example comes from Gorard (2016).

Each school’s GS residual is the absolute value of the result of subtracting the population proportion of all students in each school from the population proportion of potentially disadvantaged students (such as those eligible for FSM) in each school. The GS itself is the sum of these residuals for all schools divided by two.

More formally, GS = 0.5 \* (Σ|Fi/F − Ti/T|),  
where:

Fi is the number of FSM children in school i

Ti is the total number of children in school i

F is the total number of FSM children in the chosen area

T is the total number of children in the chosen area.

Regarding D, the D residual is the absolute value of the result of subtracting the population proportion of non-FSM pupils in each school from the population proportion of FSM pupils in each school. D itself is the sum of these residuals for all schools divided by two.

More formally, D = 0.5 \* (Σ|Fi/F − Ni/N|),  
where:  
Fi is the number of FSM children in school i

Ni is the number of non-FSM children in school i

F is the total number of FSM children in the chosen area

N is the number of non-FSM children in the chosen area.

For a full comparison of GS and D, see Gorard (2009) and for more detailed information on GS and D, see Gorard (2009; 2016).

### 5.2.3. Linkage to NPD school-level indicators

Our data linkage consists of two steps: (1) for the six main common indicators of disadvantage in schooling (as explained above), we created school segregation data for both indices; and (2) we then linked these data to the schools of the participants in the Next Steps study. These linked data are now available in Secure Lab data waves 1-8.

Using figures available from the School-Level Annual Schools Census (SLASC) in the National Pupil Database (NPD) for state secondary schools at the national level in England from 2003 to 2006, we calculated both the GS and D for *all the secondary schools* available. The NPD has long served as one of the nation’s richest data resources and it collects administrative data from all state-maintained schools in England and provides rich school-level data including the specific schools that pupils attend, their educational attainment and “a range of possible indicators of pupil disadvantage such as eligibility for free school meals and special educational needs” (Gorard, 2013, p. 113), which can facilitate analysis of the main characteristics of school segregation.

Specifically, the indicators for the GS and D indices are as follows: 1) the number of pupils taking free school meals (FSMt), 2) the number of pupils known to be eligible for free school meals (FSMe), 3) the number known to have a statement of special educational needs (SENs), 4) the number known to have special needs without a statement (SENn), 5) the number of pupils classified as non-white British ethnic origin (NW); and 6) the number known to have English as a second language (ESL).

Table 1 below summarises the number of cases (secondary schools) available each year (2003-2006)

**Table 1: Number of School Cases**

| Year | Cases |
| --- | --- |
| 2003 | 7137 |
| 2004 | 3235 |
| 2005 | 7344 |
| 2006 | 7219 |

### 5. 2.4 Linkage to Next Steps schools

For the purpose of completeness in the linked files, school segregation indices for four years were used for data linking. School segregation is a consistent phenomenon so the indices do not change drastically. It is important to point out here that all the data used for the calculations are at the school level and the school unique registration numbers (URNs) were *only* used for linkage. The linkage was independently conducted by the Centre for Longitudinal Studies (CLS), UCL, Institute of Education, which was given access to the data files with matching codes for schools. After matching no individual school or student can be identified.

For the data linkage, the dataset of segregation figures produced was merged with the Next Steps dataset using the LAESTs (LA Establishment number). Some variables were relabelled and the final dataset was de-anonymised by reducing the number of decimal digits, assigning the Next Steps anonymised school identifier and removing the original URN/LAEST.

## 5.3. Potential research areas

This newly linked dataset can be a valuable source for several potential research areas: (1) a wide range of social issues relating to the clustering effects of who goes to school with whom; and (2) the wider social impact of school intake, especially in terms of equity in education.

## 5.4. Data access

Next Steps survey and linked administrative data are managed by the Centre for Longitudinal Studies (CLS) and made publicly available via the UK Data Service (UKDS):

<https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=2000030>

In particular, linked administrative education data require access via the UKDS Secure Lab, so researchers should ensure that they have an approved account in order to work with these data:

<https://www.ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab.aspx>

To access the linked school segregation data via the UKDS Secure Lab, researchers will need to apply directly to the Centre for Longitudinal Studies (CLS) by completing the CLS Data Access application, which is available at

<https://cls.ucl.ac.uk/data-access-training/access-cls-dac/>

On approval by the CLS Data Access Committee, CLS provides a copy of the data through the researcher's UKDS Secure Lab for data analysis.

# 6. References

Anders, J. (2012). Using the Longitudinal Study of Young People in England for research into Higher Education access, Department of Quantitative Social Science, DoQSS Working Paper.

Demack, S., Platts-Fowler, D., Robinson, D., Stevens, A. and Wilson, I. (2010). Young people and community cohesion – analysis from the longitudinal study of young people in England (LSYPE). Report no. DFE-RR033. London: Department for Education.

Department for Education (2011). LSYPE User Guide to Datasets: Wave 1 to Wave 7. Retrieved from <http://doc.ukdataservice.ac.uk/doc/5545/mrdoc/pdf/5545lsype_user_guide_wave_1_to_wave_7.pdf>

Gorard, S., Taylor, C. & Fitz, J. (2003). *Schools, Markets and Choice Policies*, London: RoutledgeFalmer.

Gorard, S. (2009). Does the index of segregation matter? The composition of secondary schools in England since 1996, *British Educational Research Journal, 35*:4, 639-652, DOI: 10.1080/01411920802642389.

Gorard, S. (2013). *Research Design: Creating Robust Approaches for the Social Sciences*. London: Sage.

Gorard, S. & See, B. H. (2013). *Overcoming disadvantages in Education.* London, UK: Routledge.

Gorard, S. (2014). [The link between Academies in England, pupil outcomes and local patterns of socio-economic segregation between schools](https://www.dur.ac.uk/education/staff/profile/?mode=pdetail&id=11539&sid=11539&pdetail=89494). *Research Papers in Education* 29(3): 268-284.

Gorard, S. (2016). The complex determinants of school intake characteristics and segregation, England 1989 to 2014, *Cambridge Journal of Education, 46*:1, 131-146, DOI: 10.1080/0305764X.2015.1045446.

Gorard, S. (2018). *Education Policy: Evidence of equity and effectiveness*. Bristol: Policy Press.

Goodman, A. & Gregg, P. (2010). (eds.) *Poorer children’s educational attainment: how important are attitudes and behaviour?* York, Joseph Rowntree Foundation.

Hagger-Johnson, G., Taibjee, R., Semlyen, J., Fitchie, I., Fish, J., Meads, C. & Varney, J. (2013). Sexual orientation identity in relation to smoking history and alcohol use at age 18/19: cross-sectional associations from the Longitudinal Study of Young People in England (LSYPE). BMJ Open 2013;3: e002810. DOI:10.1136/ bmjopen-2013-002810.

Note, B. (2019). Post-16 Aspirations and Outcomes: Comparison of the LSYPE Cohorts. Centre for Vocational Education Research.

Sala, E., Burton, J. & Knies, G. (2012). Correlates of obtaining informed consent to data linkage respondent, interview and interviewer characteristics. *Sociological Methods and Research, 41(3)*, 414-39.

Siddiqui N. (2019). Using secondary data in education research. *Social Research Update*. ISSN: 1360-7898.

Siddiqui, N., Boliver,V. & Gorard, S. (2019). Reliability of Longitudinal Social Surveys of Access to Higher Education: The Case of Next Steps in England. *Social Inclusion, 7*(1), 80-89.

Williams, J. & Hagger-Johnson, G. (2017). Childhood academic ability in relation to cigarette, alcohol and cannabis use from adolescence into early adulthood: Longitudinal Study of Young People in England (LSYPE). *BMJ Open* 2017;7:e012989. doi:10.1136/bmjopen-2016- 012989.