Issue 68: Spring 2019

UNIVERSITY OF SURREY

social research Update

Using secondary data in education research

Nadia Siddiqui

nadia.siddiqui@durham.ac.uk

Dr Nadia Siddiqui is Assistant Professor at Durham University. Her research is based on understanding poverty and inequalities through population data sets and large scale surveys such as National Pupil Database (NPD, England), Higher Education Statistics (HESA), Annual Survey of Education Report data (ASER, Pakistan), Longitudinal Survey of Young People in England (LSYPE). By using these secondary data resources she investigates the indicators of disadvantage that determine children's academic attainment, well-being and happiness, and access to pathways for successful life.

- **Understanding the forms of secondary datasets**
- **Strengths and limitations of secondary data resources**
- **Linking secondary data for research**

Advances in technology have made data resources more accessible and easier for research. New technologies have enhanced the capacity to store data collected through primary fieldwork. There are also efficient means to link internet based secondary data resources, in order to increase the potential for research. Open access data initiatives such as UK Data Service (UKDS), Office for National Statistics, and Organization for Economic Cooperation and Development (OECD) provide access to a large number of datasets, generated through publicly funded research projects or administrative activities. The focus here is on UK and international comparative resources. However, most countries will have similar datasets, even if they are more limited (see also Annual Status of Education Report, ASER). In social science research there is a long tradition of using secondary data, and particularly in the field of education such resources have been extensively used in research and policy development. The following secondary data resources have been used widely in social science and education research:

1. Administrative data
   a. Information from local authorities
   b. Department for Education data for schools
   c. National Health Service data
2. Social survey and longitudinal cohort study datasets
   a. Labour Force Surveys (LFS)
   b. Office for National Statistics surveys (ONS)
   c. National Child Development Study (NCDS)
   d. British Social Attitudes survey (BSA)
   e. Millennium and British Cohort Studies (MCS & BCS)
   f. Avon Longitudinal Study of Parents and Children (ALSPAC)
   g. Understanding Society
   h. Next Steps (Longitudinal Study of Young People in England LSYPE)
   i. Young Persons Behaviour & Attitude Survey (YPBAS)
   j. Twins Early Development Study (TEDS)
3. Cognitive Assessment data
   a. Programme for International Student Assessment (PISA)
   b. Trends in International Mathematics and Science Study (TIMMS)
   c. Progress in International Reading Literacy Study (PIRLS)
   d. Annual Status of Education Report (ASER)

All of the above are examples of publicly available (usually free) secondary datasets. The open access data policy encourages the use of these datasets and recently ESRC, the UK research council, has set up a research grant fund, the Secondary Data Analysis Initiative (SDAI), to promote the use of secondary data resources.

Administrative datasets

Administrative datasets are developed as a result of official procedures and record keeping. One of the main purposes of administrative procedures is to process tasks such as the distribution of salaries and pensions, fund allocation to schools and hospitals, and human resource management. The administration and delivery of such services require meticulous record keeping which generates a large quantity of data. Some administrative datasets are also known as 'big data' because information is stored at a large scale (Connelly et al. 2016). For research, this data is valuable because it is collected at a population scale and through standard administrative procedures.

An example is the data generated by the Department for Education (DfE) in England. The DfE collects information on all children attending mainstream school in a repository called the National Pupil Database (NPD). This stores administrative records of children from the first day they enter school until they complete compulsory school education. Individual pupil level information is longitudinally linked with their academic attainment and school characteristics. Educational datasets, equivalent or similar to NPD, are available in other countries where education systems are fully or partially state-funded.

The limitations of administrative datasets are:

1. Information is meticulously recorded with high reliability checks, but it is very narrow and can only be used as an indicator of certain characteristics. For example, the Free School Meal (FSM) status of a child is recorded as 'Yes' or 'No'. This is an indicator of household poverty according to national threshold levels of minimum income required to cover basic needs, but FSM does not explain poverty in terms of level of deprivation and length of time in poverty. Not all households live in the same level of deprivation, and some would even be touching the poverty threshold level without being identified as poor (Gorard 2012 and Taylor 2018).
2. Missing data is the most challenging feature of secondary data research. Although administrative records collect data meticulously, there are missing records. For example, In case of FSM status, 6% children do not have their FSM status recorded.
3. Administrative procedures require reliability checks through proof of official documents such as passports for asylum seekers, official statement of receiving social benefits etc. In absence of such proofs, the administrative records do not enter the unverified details. It is likely that those who miss such proofs of identification are vulnerable (Gorard et al. 2017).
4. Public access to data covering sensitive characteristics and personal details is restricted. Full data access needs approval and access to a highly secure IT environment in selected institutions.

Social Surveys

Social surveys are based on samples from the population of interest. Labour Force Surveys and Attitude Surveys are based on a cross-sectional research design which means that the sample is targeted for the survey activity in one frame of time. The surveys are administered following a standard protocol in which the survey instruments, survey administration and data recording are performed in a standardised way (Lynn 2010).

Social survey administration is a time consuming and costly activity as it requires approaching and recording details of a large number of people. Increasingly, social surveys are using internet platforms to approach the targeted sample which has reduced the cost of social survey administration, but this can limit the representativeness of the sample.

Unlike administrative data, social survey data can give detailed information about participants' attitudes, feelings and experiences. The information collected in social surveys is broad and can include a range of measures that are not available in administrative records. For example, in social attitudes surveys such as BSA and YPBAS, there is information on participants' social attitudes, views and experiences such as gender roles, voting, climate change, social trust, health and wellbeing.

The limitations of social survey data are:

1. Although a range of information is available about participants, it is based on self-reports. Respondents provide their personal details which might not truly reflect their circumstances. Misreporting by respondents could be due to reasons such as lack of information, poor understanding, unwillingness to disclose information, survey fatigue etc.
2. The targeted sample participants might refuse or withdraw participation. In some social surveys participant non-response and withdrawal is not clearly reported. There is always some level of non-response and withdrawal which is unavoidable. Removing or imputing cases with missing data could give misleading

results.

Longitudinal and cohort study datasets

Longitudinal study datasets are a valuable resource for research as they allow studying patterns and trends over a period of time. The National Child Development Study (NCDS) followed a birth cohort of 17000 children born in 1958 and the last wave of data was collected in 2003 from 9000 NCDS participants. This data, collected systematically over about 45 years, is a valuable resource for research on lifelong trajectories and outcomes. The British Cohort Study (BCS) followed a cohort of 17000 individual born in a single week of 1970. There have been 8 data sweeps of this study. The Millennium Cohort Study (MCS) selected a sample of 19000 children born in 2000 and the last data sweep took place in 2017. Although these longitudinal studies are valuable, there are some limitations in the datasets:

1. At each data sweep participants drop out. There is substantial evidence which shows that participant dropout is not random. Those who are socioeconomically disadvantaged tend to drop out earlier in the study, and therefore the sample retained at the later stages is not a balanced representation of the initially selected group. For more details see Siddiqui, Boliver and Gorard (2019).
2. Missing data is a concern for the validity of the study. For example, in Next Steps (LSYPE), only 47% of the households reported their annual income. A majority of those who did were above the national income threshold of poverty. (For more details, see Siddiqui, Gorard and Boliver (2019).
3. The data is systematically collected by using highly structured and sensitive measurement tools which are administered by trained researchers. However, the information collected from the study participants is self-reported.

Cognitive assessment data

Since the late 1990s, trends in education and academic achievement have been assessed using international comparisons. This research has led to a better understanding of education policies and practices in different national contexts, and an international ranking of nations based on a standard measure of children’s cognitive abilities. These regular assessments generate datasets on children’s cognitive abilities and systematic records of a wide range of factors associated with children’s learning that are excellent resources for secondary data analysis.

Trends in International Mathematics and Science Study (TIMMS) and Progress in International Reading Literacy Study (PIRLS) are longitudinal assessments of children’s cohort sampled and assessed in Year 3-4 of primary schools and then re-assessed in the final year of secondary schools. The assessments include achievement in mathematics and science (TIMSS) and reading (PIRLS) in more than 60 countries.

The Programme for International Student Assessment (PISA) is a worldwide study of 15 years old children living in countries in the Organisation for Economic Co-operation and Development (OECD). PISA activity is repeated every three years on a sample of schools and each year new samples are selected. The assessments are based on performance in mathematics, science, and reading. In addition to the children’s assessments, additional details on parents, teacher and schools are collected. The datasets can be compared to enable countries to improve their education policies and outcomes. PISA, TIMMS and PIRLS datasets are open access and have been extensively used by the OECD countries in developing education policies.

The limitations of international assessments are that they are based on samples of schools or children. Although the sampling strategy is designed to include a wide range of groups across different national contexts, there is still a non-response from the targeted sample which is not fully reported. Missing data is another challenge which is often neglected in analyses.

Linking datasets

Some existing datasets can be linked for the purpose of research. The linking relies on common identification codes between the datasets. Pupils, schools, universities and hospitals have unique identification codes which are useful for data management and analysis. Linking datasets can increase the scope for social science research but at the same time it has various legal and ethical implications. Linked datasets are sometime only made available in a secure research environment because the individual details could be sensitive with a high risk of participant disclosure.

The most common reason to link datasets is to include a variety of indicators in the analyses. For example, Higher Education Statistical Agency (HESA) is a rich data resource on students in the UK higher education institutions. In order to understand students’ access patterns to university with regards to their prior school attainment, HESA could be linked with the NPD using pupils’ unique identifier code.

The samples of students and schools in cohort studies such as TEDS, LSYPE, BCS and MCS have been linked with administrative records of schools and the NPD. This linking allows access to additional information about participants such as their attainment records at key stages in education. Some studies have also used linked datasets to assess the reliability of indicators available from different data resources. For example a recent study compared Free School Meal (FSM) status as recorded in NPD with household Income as reported in MCS and LSYPE (Taylor 2018 and Siddiqui et al. 2019).

Another example of linking datasets is for evaluation studies sponsored by the Education Endowment Foundation (EEF) which assess the impact of educational programmes on disadvantaged pupils' attainment. These large scale evaluation studies include standardised assessments, but a number of these evaluations link pupils' attainment in national assessments to measure the impact of interventions. Some key indicators of pupils' characteristics as recorded in NPD are also used for the sub-group analysis such as the impact of programmes on pupils with special education need or English as additional language.

Linking datasets is useful for research in social sciences, but has various limitations as well. Approval to engage in linking can require time-consuming security checks. In some linked datasets, important information has to be suppressed or deleted to protect individuals' identities. LSYPE has been linked with participants' attainment at GCSE and A-level as recorded in the NPD. However, the linked data is not complete and important details have been suppressed or not made available for the participants who refused to participate and/ or dropped out of the study at a later stage. Errors in the process of linking information in large datasets can also give incomplete or misleading findings.

Combining research approaches and datasets

The combination of primary and secondary data resources can bring rigour to research by selecting the most appropriate and best measures for analysis. Good examples of such combined data approaches are the evaluation studies in education in which primary field work involved collecting data on the implementation of interventions and feedback from stakeholders (Siddiqui et al. 2018). However, for the main impact outcomes pupil level data was extracted from NPD because it is independent of stakeholders with vested interests. This combination of data resources is also good value for money because using external standardised assessments increases the financial cost of a research project, and additional resources are also required for the administration of assessments. Key stage national assessment data is available for all pupils and it is freely accessible for research use (for example evaluation studies, see Siddiqui et al. 2017 and 2018).

References

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, 59, 1-12.

Gorard, S. (2012). Who is eligible for free school meals? characterising FSM as a measure of disadvantage in England. *British Educational Research Journal,* 38(6): 1003-1017.

Gorard, S., Siddiqui, N. & Boliver, V. (2017). An analysis of school-based contextual indicators for possible use in widening participation, *Higher education studies*, 7 (2), 79-100.

Lynn, P. (2010) Developing quality standards for cross-national survey research: Five approaches. *International Journal of Social Research Methodology*, 6(4), 323-336.

Siddiqui, N., Gorard, S. & See, B. H. (2017). Can programmes like Philosophy for Children help schools to look beyond academic attainment?. *Educational Review*, 71(2), 146-165.

Siddiqui, N., Gorard, S. & See, B.H. (2018). The importance of process evaluation for randomised control trials in education. *Educational Research,* 60(3), 357-370.

Siddiqui, N., Boliver,V. and Gorard, S. (2019) Assessing the reliability of longitudinal social surveys of access to higher education: The case of the Next Steps survey in England. *Social Inclusion*, 7(1), 80-89.

Taylor, C. (2018). The Reliability of Free School Meal Eligibility as a Measure of Socio-Economic Disadvantage: Evidence from the Millennium Cohort Study in Wales, *British Journal of Educational Studies*, 66(1), 29-51.

social research UPDATE is distributed without charge on request to social researchers in the United Kingdom by the Department of Sociology at the University of Surrey as part of its commitment to supporting social research training and development.

Contributions to *social research UPDATE* that review current issues in social research and methodology in about 2,500 words are welcome. All *UPDATE* articles are peer-reviewed.

social research UPDATE (ISSN: 1360-7898) is published by the Department of Sociology, University of Surrey
Guildford GU2 7XH, United Kingdom. tel: +44 (0)1483689450
Edited by Nigel Gilbert (n.gilbert@surrey.ac.uk)