Article:
Cave, Sophie Nicole and von Stumm, Sophie (2020) Secondary data analysis of British population cohort studies: A practical guide for education researchers. British Journal of Educational Psychology. e12386. ISSN: 0007-0998

https://doi.org/10.1111/bjep.12386

the british psychological society
promoting excellence in psychology

British Journal of Educational Psychology (2020)

www.wileyonlinelibrary.com

Secondary data analysis of British population cohort studies: A practical guide for education researchers

Sophie Nicole Cave and Sophie von Stumm*
Department of Education, University of York, UK

Background. Britain is rich in longitudinal population cohort studies that posit valuable data resources for social science. However, education researchers currently underutilize these resources.

Aims. The current paper (1) outlines the power and benefits of secondary data analyses for educational science and (2) provides a practical guide for education researchers on the characteristics, data, and accessibility of British population cohort studies.

Methods. We identified eight British population cohort studies from the past 40 years that collected scholastic performance data during primary and secondary schooling, including (1) Avon Longitudinal Study of Parents And Children (ALSPAC), (2) Twins Early Development Study (TEDS), (3) Effective Pre-School, Primary and Secondary Education Project (EPPSE), (4) Millennium Cohort Study (MCS), (5) Born in Bradford (BiB), (6) Next Steps (LYSPE1), (7) Understanding Society (US), and (8) Our Future (LYSPE2). Participants across these studies were born between 1989 and 2010, and followed up at least once and up to 68 times, over periods of 7 to 29 years. For each study, we summarize here the context and aims, review the assessed variables, and describe the process for accessing the data.

Conclusions. We hope this article will encourage and support education researchers to widely utilize existing population cohort studies to further advance education science in Britain and elsewhere.

Population cohort studies are characterized by the year or decade of the cohort members' birth and by the geographical sampling area from which they were recruited. Population cohort studies are often observed longitudinally, with their cohort members being followed up repeatedly across the lifespan. Over the past 50 years, the Medical Research Council (MRC) has invested almost £30 million a year in a bid to support 34 of the United Kingdom's (UK) largest population cohort studies (Pell, 2014). Likewise, the Economic & Social Research Council (ESRC) spends approximately 10% of their annual budget on UK population cohort studies (Davis-Kean et al., 2017), while the Wellcome Trust has invested £120 million in UK population cohort studies as well as in those from low- and middle-income countries (Wellcome's Longitudinal Population Studies Working Group, 2017). As a result, Britain is particularly rich in nationally representative, longitudinal

**Correspondence should be addressed to Sophie von Stumm, Department of Education, University of York, YO10 5DD, Heslington, York, UK (email: sophie.vonstumm@york.ac.uk).*

DOI:10.1111/bjep.12386

population cohort studies, whose data are extensively analysed by national and international researchers across social science research disciplines, for example sociology, economics, epidemiology, genetics, and psychology. However, education scientists appear to utilize these data resources less often (Siddiqui, 2019); for example, in the three most recent issues of BJEP, none of the 32 published articles applied secondary data analyses to one of the British population cohort studies that we review here.

We speculate that education researchers may not be fully aware of the advantages of secondary data analysis and how to best access and utilize the available population cohort studies, because the latter are not routinely covered in undergraduate and postgraduate training in education science. To promote the broader use of population cohort studies, we outline first the power and benefits of secondary data analysis for advancing educational science, and we then review the British population cohort studies that emerged during the past 40 years and assessed, among other variables, children's performance during primary and secondary school. Studying school performance is imperative for education researchers, because it serves two important functions. For one, school performance indicates the extent to which children have mastered the knowledge and skills that are essential for them to successfully participate in society, for example, reading, writing, and arithmetic. For the other, school performance functions as a gatekeeper that regulates children's access to further education (Danilowicz-Gösele, Lerche, Meya, & Schwager, 2017; von Stumm et al., 2020). School performance pertains to a plethora of research topics, ranging from – but not limited to – understanding the genetic and environmental factors that influence children's differences in learning ability (Krapohl & Plomin, 2016), to studying the role of personality traits for how children learn and retain information (Komarraju, Karau, Schmeck, & Avdic, 2011; Vaishnav & Chirayu, 2013), and to exploring gender differences in educational achievement (Matthews, Ponitz, & Morrison, 2009; Weis, Heikamp, & Trommsdorff, 2013). With this article, we aim to encourage educational scientists to enrich their programmes of research by leveraging the population cohort studies that are high-quality data resources available in Britain.

Strengths of secondary data analysis

Original or primary data collection is extremely costly in time and effort (Queirós, Faria, & Almeida, 2017). As a result, samples obtained through original or primary data collection are often modest in size, which makes them susceptible to biases from non-representative sampling and incomplete data (Cheema, 2014; Davis-Kean & Jager, 2012). Secondary data analyses of population cohort studies overcome these limitations, because they rely on large samples that have been broadly assessed using state-of-the-art measures. It follows that population cohort studies enable well-powered studies of high scientific rigour and validity, whose findings generalize widely (Davis-Kean & Jager, 2012; Davis-Kean, Jager, & Maslowsky, 2015; Smith et al., 2011), although they are often affected by attrition, which can cause sampling biases (Duncan & Gibson-Davis, 2006; Watson & Wooden, 2009). The scientific power of secondary data analyses can be further improved when researchers engage in cross-cohort collaborations (Pell, Valentine, & Inskip, 2014), harmonize data across samples, and conduct data linkage across data repositories (Jay, Mc Grath-Lone, & Gilbert, 2019).

Securing the funding for original or primary data collection, including the recruitment, assessment, and compensation of participants, can take many years, as does the coding,

Table 1. British population cohort studies from the past 40 years that collected school performance data

Cohort acronym	Scope	Year	$N_{Recruitment}$	$N_{Education}$	$\%_{Educatio}$	$Age_{Education}$	n	EYFSP	KS1	KS2	KS3	KS4	Access	$Security_{Level}$	Fee
ALSPAC	County - Avon	1991–1992	14,500	11,300	5–7		78		*X*	*X*	*X*	*X*	Direct		*X*
TEDS	England & Wales	1994–1996	26,000	12,500	7		48			*X*	*X*	*X*	Direct		
EPPSE[a]	UK Wide	1997	3,200	3,200	5–7		100		*X*	*X*	*X*	*X*	UKDS	Safeguarded	
MCS	UK Wide	2000–2001	19,000	11,900	5		63	*X*	*X*	*X*		*X*	UKDS	Safeguarded & Controlled	
LYSPE1[a]	England	2004	15,770	14,800	8–11		94			*X*	*X*	*X*	UKDS	Controlled	
BiB	City – Bradford	2007–2010	14,000	10,600	5		76	*X*	*X*				Direct		*X*
US[b]	UK Wide	2009–2011	51,000	2,000	5		4[c]	*X*	*X*	*X*	*X*	*X*	UKDS	Controlled	
LYSPE2[a]	England	2013	13,000	12,200	5–7		93		*X*			*X*	UKDS	Controlled	

Note. Cohort acronym refers to the abbreviated cohort names. Scope refers to the cohort's geographical sampling area. Year refers to year of birth, except for cohorts where birth years differed from year of the study start; in these cases, year of study start is shown[a]. $N_{Recruitment}$ refers to the total number of participants at wave one. $N_{Education}$ refers to the number of children whom education data is available for at the earliest assessment age (i.e. $Age_{Education}$). $\%_{Education}$ is the proportion of the sample with education data, relative to $N_{Recruitment}$. EYFSP refers to Early Years Foundations Profile Scores; KS1–KS4 refer to Key Stages 1 through 4. [b]US includes a relatively small proportion of households with school-aged children for whom school performance data are available. Access refers to whether an application for data usage is made through the UK Data Service (UKDS) or directly through the cohort steering committee. $Security_{Level}$ applies to datasets held by the UK Data Service. Fee refers to any associated finances required to obtain the data.

cleaning, and archiving of data for analysis. By comparison, the population cohort studies we describe here are far less expensive to utilize, and most of them can be accessed quickly and free of charge. Even when population cohort studies require payment of access fees, they are a fraction the costs of primary or original data collection. Secondary data analyses of population cohort studies are therefore highly cost-effective (Johnston, 2017; Smith, 2008), which makes them an appealing resource for researchers at all stages of their careers, who wish to build their academic portfolios (Hakim, 1982).

In Britain, specific funding schemes have been designed to support researchers who seek to conduct secondary data analysis (e.g. ESRC, SDAI: https://www.ukri.org/opportunity/secondary-data-analysis-initiative). This is a notable exception to the priorities of funding agencies in other countries that often accept secondary data analyses as sustainable research method but prioritize original or primary data collection.

Secondary data analyses of existing population cohort studies offer a great number of opportunities for novel empirical discoveries, as well as for replications and extensions of previous findings (Andrews, Higgins, Andrews, & Lalor, 2012; Davis-Kean et al., 2015). For example, researchers have utilized these rich resources to explore school performance in relation to child poverty and mental health (Nikulina, Widom, & Czaja, 2011), physical activity (Donnelly et al., 2016), and attention difficulties (Polderman, Boomsma, Bartels, Verhulst, & Huizink, 2010). Indeed, population cohort studies offer much more data than a single researcher could collect; these data make innovative and original research possible.

Population cohort studies can also support education researchers in exploring societal, historical and governmental trends over time (Jay et al., 2019). For example, secondary analyses of the population cohort studies described here can serve to explore whether and how changes to the British education system are reflected by students' achievements. For example, the Pupil Premium, which was introduced in 2011 by the British coalition government under David Cameron (2010–2015), is a grant awarded to schools that enrol pupils from impoverished and unstable family homes to fund educational resources for these pupils to overcome their disadvantages. The effectiveness of this policy for reducing the influence of family background on school performance could be established through comparisons of pupil populations that attended school before and after the Pupil Premium was brought in (Lupton et al., 2015). However, no population cohort study has been conceived since the advent of the Pupil Premium and thus, school performance data from a cohort that experienced this policy are not available. Typically, population cohort studies are not created in response to or for tracking policy changes, limiting their suitability for analysing the effectiveness of interventions (Duncan & Gibson-Davis, 2006).

School performance in Britain

We focused on scholastic performance during primary and secondary school due to its relevance to education researchers (see above) and because of its pivotal role for people's life outcomes (Schoon, Jones, Cheng, & Maughan, 2012; von Stumm et al., 2020). In Britain, school performance is captured through four statutory Key Stage (KS) assessments that are completed at children's ages 7, 11, 14, and 16 years (i.e. KS1, KS2, KS3, and KS4). In addition, the Early Years Foundation Stage Profile (EYFSP), introduced in 2008, rates children's knowledge and progress at the end of reception (aged 5 years). First

defined under the Education Reform Act of 1988, KS tests now assess children's understanding of the National Curriculum in England and Wales (Hutchison & Schagen, 1994). Because the National Curriculum applies to all local authority-maintained schools, KS grades can be directly compared across institutions, regions, and time.

We note that the National Curriculum differs across UK countries (i.e. England, Wales, Scotland, and Northern Ireland). A decade ago, Scotland introduced the 2010 Curriculum for Excellence to guide children's development, while Northern Ireland announced in 2007 their own country's curriculum to cover all 12 years of compulsory education. Scotland's and Northern Ireland's curricula are not exact matches to the National Curriculum in England and Wales.

In 2012, the Department for Education (DfE) founded the National Pupil Database (NPD) for England and Wales to record students' exam results, teacher reported predicted grades, and how many qualifications students achieved (Jay et al., 2019). The NPD is an extensive resource which also pertains information on school demographics, attendance, and additional support students may require (https://find-npd-data.education.gov.uk). Since its inception, students' NPD records have been linked, where possible, with their data that were collected through population cohort studies. As a result, many UK population cohort studies are now enriched with longitudinal scholastic school performance data, making them particularly valuable resources for education researchers.

The current article

At present, no practical guide exists for education researchers about the British population cohort studies that are available for secondary data analysis, although Siddiqui (2019) wrote an excellent general introduction to the topic. By contrast, for health scientists and epidemiologists, the MRC has produced a strategic review of 34 population cohorts in the United Kingdom (Pell, 2014). Similarly, for government employees, the DfE produced a document summarizing longitudinal surveys on children and young people under the age of 19 years (DfE,). Although these are valuable

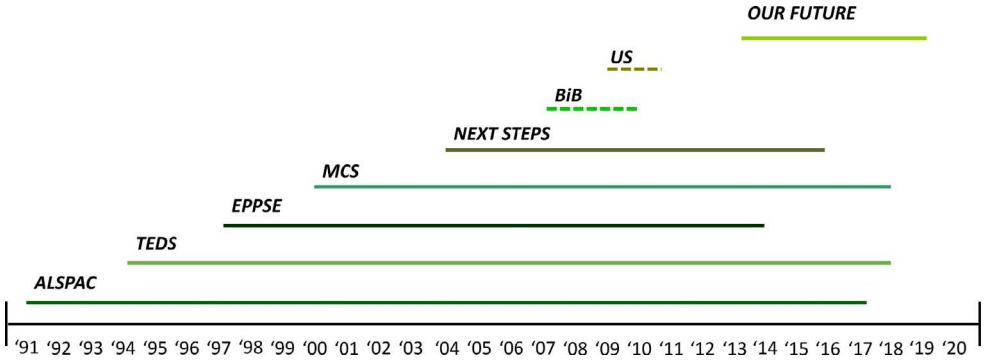

Figure 1. British population cohort studies by their period of observation.

Note: Studies are plotted from the year of inception through to year of last assessment. Straight lines refer to population cohort studies whose members have been repeatedly followed up and assessed. Dashed lines are for studies that conducted one assessment wave and follow-up data are available for subsamples or from linkage with the National Pupil Database (NPD) and other national data repositories.

resources, they do not focus on school performance data and are of limited utility for education researchers.

Here we describe eight cohort studies that sample populations born in Britain during the past 40 years. We identified them through searching online repository archives, such as CLOSER (www.closer.ac.uk) and the Centre for Longitudinal Studies (https://cls.ucl.ac.uk), through published cohort profiles, and through consultations with experts in longitudinal data analysis. We followed criteria similar to those that informed the MRC and DfE reports to select the population cohort studies for our review: (1) to be longitudinal or cohort in nature; (2) draw their sample from an area in the United Kingdom that is broadly representative of Britain; (3) to have at least 1,000 participants upon first recruitment; (4) to be conducted within the last 40 years; and (5) to include validated measures of school performance data collected during primary and/or secondary school. We excluded studies from our review whose school performance data were collected but are not available for researchers (e.g. Growing Up in Scotland; Anderson et al., 2007).

The population cohort studies

We identified eight population cohorts: the Avon Longitudinal Study of Parents And Children (ALSPAC); the Twins Early Development Study (TEDS); the Effective Pre-School, Primary and Secondary Education Project (EPPSE); the Millennium Cohort Study (MCS); Next Steps (LYSPE1); Born in Bradford (BiB); Understanding Society (US); and Our Future (LYSPE2). At the population cohort studies' inception, their sample sizes ranged from 3,000 participants to over 51,000 (Table 1). For most of the population cohorts, school performance is only available for a proportion of the initial sample, with some collecting or linking school performance data at multiple time points, and others only once. As typical in longitudinal research, the population cohort studies identified here are all affected, albeit to different agrees, by missing data due to attrition, which can cause sample biases (Watson & Wooden, 2009; see Table 1). A plethora of methods are available to researchers to deal with missing data due to attrition, for example applying sampling weights, imputation, and using appropriate statistical estimators (e.g. full-information maximum likelihood; cf. Duncan et al., 2003; Duncan & Gibson-Davis, 2006). A review of these is beyond the scope of the current work, but interested readers may consult the respective cohorts' user guides and principle statistics texts on this topic for further guidance. Figure 1 plots the identified UK population cohort studies by their period of observation.

Across the population cohorts identified here, the age of assessment of school performance ranged from the start of primary school (age 5 years, reception class) to the end of compulsory school (i.e. age 16 years until 2015, then raised to age 18 years; Figure 1). In the current article, we only focus on population cohorts' assessment of school performance up to age 16, although some studies also collected performance data at later education stages. However, reviewing the education data that are available in population cohort studies post age 16, when education trajectories become increasingly varied, is beyond the scope of the current paper. Further details on the population cohorts' education data post age 16 can be found within the respective verbose codebooks, data dictionaries, and cohort profiles (e.g. Clark, Demster, & Solberg, 2012; Ferrie, 2012; McCabe et al., 2019).

Below we briefly describe each identified population cohort and summarize their school performance measures. We also explain the studies' access procedures for researchers who seek to engage in secondary data analysis that is in the public interest and is not being carried out for personal or commercial gain. Researchers who require data

from population cohort studies for other purposes are advised to contact the respective study's steering committee. To ensure safe use of data, researchers must abide by General Data Protection Regulations (GDPR; Information Commissioner's Office, 2018) when applying for, accessing, and analysing data. For queries regarding GDPR, researchers must contact the respective cohort's Data Protection Team, who oversees the data application and approval process. Before describing each population cohort, we review the difference between 'safeguarded' and 'controlled' data and how it affects data access.

Accessing 'safeguarded' and 'controlled' data

Established in 2012 with funding from the ESRC, the UK Data Service currently holds over 7,000 digital data collections. For population cohort studies stored within the UK Data Service, access arrangements are dependent on whether data owners have classified the datasets as 'safeguarded' or 'controlled'. 'Safeguarded' data is provided under the UK Data Service's End User Licence (https://www.ukdataservice.ac.uk/get-data/how-to-access/conditions.aspx), which implies that although the data are not personal, the data owners have classified them as potentially disclosive when linked to other databases. To access 'safeguarded' data, researchers must register with the UK Data Service and accept the End User Licence agreement. This process takes less than an hour and grants researchers immediate data access (https://www.ukdataservice.ac.uk/get-data/data-access-policy.aspx).

'Controlled' data includes personal data that make individuals identifiable, and thus, these data are potentially disclosive. Data classified as 'controlled' cannot be downloaded and accessed directly by researchers. In general, controlled data can only be accessed (1) from a UK location, (2) by researchers with a UK Higher Education or Further Education affiliation, (3) if data are not going to be used for commercial purposes, and (d) if data use has been approved by the data owners (e.g. Professor Emla Fitzsimons as Director of the MCS at the time of writing). If these conditions are met, researchers must complete the following steps to access 'controlled' data: (1) read and accept the End User Licence agreement; (2) fill in and return the ESRC Accredited Researcher Proposal to outline in principle the scope of the planned research, including variables required, statistical analysis, and the implications of the findings; (3) fill in and return the ESRC Accredited Research Application, which details contributions to journals and technical access arrangements; and (4) fill in and return the Secure Access User Agreement, to be completed by each person who will have access to the data and signed by the Principal Investigator and his or her host institution's legal team. Once completed, the ESRC Accredited Researcher Proposal, Research Application and Secure Access User Agreement are to be returned via email to secure.applications@ukdataservice.ac.uk.

First-time applicants for 'controlled' data access qualify as a new researchers, and they have to complete Safe Researcher training course in addition (researchers who have completed the course since 1st January 2016 will have to complete a short online refresher) . At the time of writing, the Safe Reseacher training is delivered in person during workshops that take place in London or Colchester. The one-day training course is based on the Five Safe's (Desai, Ritchie, & Welpton, 2016), a security model which ensures data, projects, people, settings and outputs are safe, and introduces users to the UK Data Service and the Secure Lab.

A researcher will be granted data access once their request has been approved and a Secure Lab account has been created. Depending on how restrictive and sensitive the data is, researchers will be able to access data either remotely, on their institution's desktop

computer via a secure virtual private network, or physically by attend the UK Data Service's safe room located at the University of Essex. Accessing 'controlled' data is a considerably lengthy process that takes approximately 9 months at the time of writing. A useful guide to support researchers through the application process has been written by Corti, Van den Eynden, Bishop, and Woollard (2019).

Avon Longitudinal Study of Parents and Children (ALSPAC)

The ALSPAC followed all pregnant women in the county of Avon, whose estimated delivery date fell between 1st April 1991 and 31st December 1992, inclusive (http://www.bristol.ac.uk/alspac). The cohort study, core funded by the MRC, Wellcome Trust, and the University of Bristol, included an initial sample of 14,541 pregnancies that resulted in 14,062 live births.

The study children were followed over the course of their development, with 78% of them having school performance data linked from the NPD, including KS1 Reading, Writing, Spelling and Maths; KS2 and KS3 Maths, English and Science; and KS4 General Certificate of Secondary Education (GCSE) and Business and Technology Education Council (BTEC) results. In addition to school performance data, ALSPAC assessed an extensive variety of measures, including but not limited to, clinical information on physical development, parents' attitudes and expectation of the child, as well as biological samples and information from mothers and partners about development and family background. A comprehensive list of all measures is available on the study's website (http://www.bristol.ac.uk/alspac/researchers), alongside the cohort profiles (Boyd et al., 2013; Northstone et al., 2019).

To access ALSPAC data, researchers must apply by completing an online proposal form, with the outcome typically communicated within two weeks. An access fee is calculated on a project-by-project basis, depending on the funding status and complexity of the project, as well as on the type of variables requested. As of May 2020, access fees for ALSPAC started from £2,105 (all figures excluding Value Added Tax), with additional charges for the extraction and inclusion of education data of approximately £1,000. If researchers propose secondary data analyses of ALSPAC in a funding bid, they are asked to include a data management fee of £7,500 to cover all data related costs. Also, they must complete the online proposal form for the ALSPAC access application at least one month prior to the funding bid's submission deadline. Further details can be found here: http://www.bristol.ac.uk/media-library/sites/alspac/documents/researchers/data-access/ALSPAC_Access_Policy.pdf. At a first glance, ALSPAC's access fee charges may seem steep for individual investigators, especially for early career researchers, who typically have little funding available to them. However, the fees are essential to maintain and to continue to collect ALSPAC's diverse and rich data at the highest scientific standards. For researchers who cannot afford ALSPAC, we recommend focusing on the population cohort studies that are more affordable or free of charge, which we describe below.

Twins Early Development Study (TEDS)

Parents of all twins born in England and Wales from 1994 to 1996 were contacted to take part in the TEDS (www.teds.ac.uk). The study aims to explore how genetic and the environmental factors influenced individual differences in affect, behaviour, and cognition. The project has been continually funded by the MRC and is based at King's College London. Over 13,000 families (i.e. 26,000 children) participated in the first

assessment wave, and they have since been followed up every two to three years until most recently at age 22 years (Oliver & Plomin, 2007; Rimfeld et al., 2019). The TEDS twins have been comprehensively assessed on a broad range of measures, including their early life experiences, cognitive and social-emotional development, learning competencies, and mental health.

Parent-, teacher-, and child-reported school performance data are available for about 48% of the sample. Teachers completed National Curriculum rating scores for English and Maths when the children were aged 7, then additionally for Science at ages 9, 10, and 12. KS3 data was provided by parents when the twins were aged 14 years, while GCSE and other examination results achieved by the students at aged 16 (KS4) were provided by the study participants themselves.

To access the TEDS data, researchers must contact the core member of the TEDS research team whose interests are best aligned with the researcher's planned project. A list of team members can be found on the data request form (https://www.teds.ac.uk/researchers/teds-data-access-policy). Researchers then complete an online data access application form with support from the TEDS core member, with the outcome being communicated within two weeks of the application submission. If data approval is granted, researchers must pre-register their study on the Open Science Framework (OSF; https://help.osf.io/hc/en-us), before the data are released. In addition, data sharing agreements need to be in place between King's College London, where the TEDS is hosted, and the applying researchers' host institution; this process takes at least 3 months.

The Effective Pre-School, Primary and Secondary Education Project (EPPSE)

Established in 1997, The EPPSE Project aimed to explore the impact of early year's education across development (https://www.ucl.ac.uk/ioe/research-projects/2020/sep/effective-pre-school-primary-and-secondary-education-project-eppse). Over 3,000 children were tracked from the start of pre-school, at 3 years old, through primary school at the ages of 6, 7, 10, and 11 years, and during secondary education at ages 14 and 16 years (Taggart, Sylva, Melhuish, Sammons, & Siraj, 2015). The project was funded by the DfE and ran from 1997 to 2013, with no further assessment waves currently planned.

School performance data for all EPPSE members were extracted from the NPD. National Curriculum ratings are available for children in Year 2 (KS1), for English (speaking & listening; reading; writing), Maths (using & applying; number & algebra; shape, space & measures) and Science (experimental & investigative/scientific enquiry; life process & living things; material & properties; physical processes). The results of KS2 Statutory Assessment Tests (SATS) are available for children in Year 6. English, Maths and Science scores are available for Year 9 students (KS3), while GCSE results are available for Year 11 students (KS4). In addition, Year 1 and Year 5 children also sat National Foundation for Educational Research (NFER) Tests in primary Reading and Maths. Researchers may also utilize the study's extensive data on child care settings, developmental problems and illnesses, and family composition.

The EPPSE data is classified as 'safeguarded' (see above) and can be accessed via the UK Data Service (Study Number 'SN 7540'; https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7540).

Millennium Cohort Study (MCS)

The MCS, also known as 'Child of the New Century', includes 18,818 infants born in 2000 and 2001, across all countries of the United Kingdom (https://cls.ucl.ac.uk/cls-studies/

millennium-cohort-study/). The study, core funded by the ESRC, serves to explore a range of topics including, but not limited to, child development, physical health, and social-emotional well-being (Connelly & Platt, 2014).

For about 63% of MCS children, school performance data are available through linkage with the NPD. KS1, KS2, and KS4 pupil level linked data exists for cohort members in England, while in Scotland and Wales, KS1 data are also available, with plans to link KS2. In addition to these statutory assessments, teachers were also asked to complete the Early Years Foundation Stage Profile (EYFSP), a legislative profile which summarizes and describes a child's attainment at the end of reception (aged 5). At the time of the survey, the EYFSP was compulsory in England but teachers in Wales, Scotland, and Northern Ireland did not complete comparable assessments. Thus, a 16-page teacher survey was specifically designed to mimic the EYFSP and administered to all teachers of MCS children across UK countries. The cohort members have been repeatedly assessed on a comprehensive range of measures, including their early life experiences and pre-school education, physical and cognitive development, and experiences of bulling and antisocial behaviour.

Access to the MCS data is via the UK Data Service. The teacher survey and foundation stage profile dataset 'SN 6847' is classified as 'safeguarded' (https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=6847). The linked education administrative datasets for England (NPD – SN 8481; https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8481), Wales (KS1 – SN 7415; https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7415), and Scotland (KS1 – SN 7414; https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7414) are classified as 'controlled'. A detailed description of the access requirements pertaining to 'safeguarded' and 'controlled' data are at the start of this section.

Next Steps (LYSPE1)

Early labour market experiences and educational prospects of young people are a key focus of the Next Steps study (https://cls.ucl.ac.uk/cls-studies/next-steps/). Funded and managed by the DfE from 2004 to 2010, the project, also known as the Longitudinal Study of Young People in England (LSYPE1), followed the lives of 15,770 individuals born in 1989 and 1990 (DfE,). The study aims to map students' educational journeys from school, to higher education and into the workplace. The cohort members have been regularly assessed on a wide range of measures, including attitudes towards education, aspirations and expectations, antisocial behaviours, health and well-being, and family formation. In 2013, the Centre for Longitudinal Studies (CLS) took over the management of the study and commissioned further exploratory work into the cohort members employment, housing, and financial situations at age 25. There are plans to conduct another assessment wave in 2021–2022.

For 94% of the LYSPE1 cohort members school performance data were linked from the NPD. This includes KS2 and KS3 Maths, English and Science; and KS4 GCSE results. In addition to school performance data, LYSPE1 also has information on free school meal eligibility and Special Education Needs & Disability (SEND) status.

The LYSPE1 is classified as 'controlled' and can be accessed via the UK Data Service (details above; Study Number 'SN 7140'; https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7104).

Born in Bradford (BiB)

A total of 13,858 children, born between March 2007 and December 2010 at Bradford Royal Infirmary, were recruited as part of the BiB study (https://borninbradford.nhs.uk; Wright et al., 2013). The project, commissioned by the Programme Grants for Applied Research funding scheme and National Institute for Health (NIH) Research Collaboration for Applied Health Research and Care, explores well-being, genetics and family environments. All children were assessed by health workers at 2 weeks, 7 weeks, and 8 months old. Several sub-studies evolved in conjunction with different funding bodies, including for example Born in Bradford's Better Start (BiBBS; Dickerson et al., 2016).

For 76% of the BiB children, school performance data was linked from the NPD at ages 5, 6, and 7 years. These data include the teacher-led EYFSP for children in Reception; a teacher-administered Phonics Assessment completed by children at the end of Year 1; and KS1 Statutory Assessment Tests in English, Maths, Science, Reading and Writing for children in Year 2. Other BiB data include Local Authority information on children's eligibility for free school meals and SEND, as well as additional demographic information collected from mothers and fathers on health, family environments and diet.

To access BiB, researchers must read the Guidance for Collaboration document on the BiB website (https://borninbradford.nhs.uk/research/how-to-access-data/). The document outlines conditions of use, as well as pertaining information on how to access biological samples. Researchers complete an online 'Expression of Interest Proforma' to describe their planned research, required variables, and statistical analysis. The 'Expression of Interest Proforma' are reviewed by the BiB Executive Group on a monthly basis. Once approved, a collaboration agreement is signed by the researcher, the BiB team, and their respective institutions' legal departments. BiB charges a data access fee of £1,000.

Understanding Society (US)

US was established in 2009 and is the continuation of the British Household Panel Survey that was conducted from 1991 to 2008 (Buck & McFall, 2011). Hosted by the Institute for Social and Economic Research at the University of Essex, US aims to track economic and social change in Britain through the collection of individual and household-level data (https://www.understandingsociety.ac.uk). US collected information from approximately 40,000 households across countries of the United Kingdom.

A small proportion of the sample are children (4%), for whom school performance data is available. For those in reception, the EYFSP has been linked from the NPD, while KS1 through to KS4 pupil level National Curriculum results are available for children in England and Wales. Subjects include English, Maths, and Science, as well as information on GCSE and BTEC results, and school absences and exclusions. In addition, parents and carers have been extensively and broadly assessed on, including but not limited to, parenting styles, family networks, and employment.

US data is classified as 'controlled' and can be accessed via the UK Data Service (see details above; Study Number 'SN 7642'; https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7642).

Our Future (LYSPE2)

Established in 2013, Our Future is the second Longitudinal Study of Young People in England (LSYPE2; https://www.ourfuturestudy.co.uk). Funded and commissioned by the DfE, LYSPE2 has followed the lives of 13,000 young people. The cohort members were interviewed yearly between the ages of 13 and 20, with the aim to explore pupils' transitions from compulsory schooling to tertiary education. The cohort members were assessed on their higher education choices, careers aspirations, employment opportunities, and health and well-being.

For 93% of the LSYPE2 cohort, school performance data have been linked from the NPD. This includes KS1 National Curriculum scores for Speaking & Listening, Reading & Writing, Mathematics, and Science, as well as KS4 GCSE and equivalent results. In addition to school performance data, LYSPE2 also has school census data including institutional type, SEND, and students' progress between KSs.

The LYSPE2 is classified as 'controlled' and accessible via the UK Data Service (details above; Study Number 'SN 7838'; https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7813).

Doing research with population cohort studies

To help planning research that builds on secondary data analyses of British population cohort studies, we describe here some of the principal demands in time, effort, and financial support that researchers need to be mindful of. Although we believe that these demands are small in comparison to those that original data collections place on the researcher, we acknowledge that they are not trivial. In general, secondary data analyses of population cohort studies typically require high-level statistical skills (e.g. Duncan et al., 2003), in part because observations are nested across levels (e.g. within families, schools, communities), and in part because of biases due to attrition and selection.

An obvious financial cost for researchers planning secondary data analysis projects are the data access fees that some population cohort studies charge. Although these are modest relative to the immense scientific value of the data, securing funding to cover data access fees can be challenging for individual investigators. As illustrated by our review above, such charges can be circumvented, however, because multiple British population cohort studies offer their data freely to researchers.

A substantial demand on time, albeit initially less apparent, stems from the extensive array of paperwork that most population cohort studies require to be completed before granting data access. The process of filling in the necessary documentation and data sharing agreements typically involves extensive collaboration and exchanges between the research team, the respective universities' legal contracts teams, and the data holders. As a result, data from some population cohorts tend to only become accessible after several months; however, others can be readily accessed at the click of a button (e.g. MCS).

Another demand in time and in effort is the training that is essential for accessing the population cohort studies. While the MRC and the National Healthcare Service run online data security courses, the UK Data Service requires all researcher to complete and pass a mandatory day of face-to-face training to access 'controlled' datasets. The training is typically offered free of charge in London or Colchester; however, attending either location for training typically requires funding to cover the costs for travel, accommodation, and expenses.

Finally, school performance data, much like health-related data, is sensitive and thus, managed through secure technical systems. While data from some population cohort studies can be downloaded directly onto personal computers (e.g. EPPSE), others require researchers to conduct their statistical analyses via monitored remote desktops (e.g. US). In some cases, researchers will be required to physically attend approved safe settings (i.e. designated office spaces), where analyses can be conducted but analysis outputs must be additionally reviewed and approved before they can be extracted from the safe setting (e.g. LYSPE2).

In summary, researchers should consider the population cohort studies' differences in their data access requirements and how they might affect the planning of research that utilizes secondary data analysis. Most of the demands and costs associated with accessing data from population cohort studies can be circumvented or managed by carefully selecting appropriate data sources. The greatest difficulty that education researchers may face when accessing data are likely to pertain to 'controlled' data that are held by the UK Data Service. Finding the right balance between ensuring the safety of personally identifiable information and making data sufficiently accessible to enable timely, impactful research is one of the greatest concurrent challenges for policy makers in education and health.

Conclusion

In the United Kingdom, one in every 30 people volunteers to contribute to a population cohort study, often without any compensation for their time, effort, and information (Pell et al., 2014). Secondary data analyses of these national data treasures offer exceptionally high value for research that generalizes to and, thus, benefits the wider public. Here, we provided a practical guide to population cohort studies that collected school performance data, with the aim to encourage education researchers to implement more often secondary data analyses of population cohort studies in their programmes of research.

Acknowledgements

We thank Katrina d'Apice, Radhika Kandaswamy, Jelena O'Reilly, Megan Wright and Allie Nancarrow for their helpful comments on an earlier draft of this paper. SvS is supported by grants from the Jacobs Foundation and the Nuffield Foundation (EDO/44110).

Conflicts of interest

All authors declare no conflict of interest.

Author contributions

Sophie Nicole Cave (Conceptualization; Investigation; Project administration; Visualization; Writing – original draft; Writing – review & editing) Sophie von Stumm (Conceptualization; Funding acquisition; Investigation; Project administration; Supervision; Validation; Writing – review & editing)

Data availability statement

No data are statistically analysed in this article.

References

Anderson, S., Bradshaw, P., Cunningham-Burley, S., Hayes, F., Jamieson, L., MacGregor, A., Wasoff, F. (2007). Growing up in Scotland: A study following the lives of Scotland's children. Scottish Executive. Retrieved from: https://era.ed.ac.uk/bitstream/handle/1842/3001/0044329.pdf?sequence=1

Andrews, L., Higgins, A., Andrews, M. W., & Lalor, J. G. (2012). Classic grounded theory to analyse secondary data: Reality and reflections. *Grounded Theory Review*, *11*(1), 12–26.

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., . . . Davey Smith, G. (2013). Cohort profile: The 'children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology*, *42*(1), 111–127. https://doi.org/10.1093/ije/dys064

Buck, N., & McFall, S. (2011). Understanding Society: Design overview. *Longitudinal and Life Course Studies*, *3*(1), 5–17. http://dx.doi.org/10.14301/llcs.v3i1.159

Cheema, J. R. (2014). Some general guidelines for choosing missing data handling methods in educational research. *Journal of Modern Applied Statistical Methods*, *13*, 53–75. https://doi.org/10.22237/jmasm/1414814520

Clark, J., Demster, B., & Solberg, C. J. (2012). Managing a data dictionary. *Journal of AHIMA*, *83*(1), 48–52.

Connelly, R., & Platt, L. (2014). Cohort profile: UK millennium Cohort study (MCS). *International Journal of Epidemiology*, *43*, 1719–1725. https://doi.org/10.1093/ije/dyu001

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2019). *Managing and sharing research data: A guide to good practice*. Thousand Oaks, CA: SAGE Publications Limited.

Danilowicz-Gösele, K., Lerche, K., Meya, J., & Schwager, R. (2017). Determinants of students' success at university. *Education Economics*, *25*, 513–532. https://doi.org/10.1080/09645292.2017.1305329

Davis-Kean, P., Chambers, R. L., Davidson, L. L., Kleinert, C., Ren, Q., & Tang, S. (2017). Longitudinal studies strategic review: 2017 Report to the Economic and Social Research Council.

Davis-Kean, P. E., & Jager, J. (2012). The use of large-scale data sets for the study of developmental science. *Handbook of Developmental Research Methods*, 148–162.

Davis-Kean, P. E., Jager, J., & Maslowsky, J. (2015). Answering developmental questions using secondary data. *Child Development Perspectives*, *9*, 256–261. https://dx.doi.org/10.1111%2Fcdep.12151

Desai, T., Ritchie, F., & Welpton, R. (2016). *Five Safes: Designing data access for research [PDF file]*. UWE Bristol Working Paper, Economics Working Paper Series 1601. Retrieved from https://uwe-repository.worktribe.com/preview/914753/1601.pdf

Dickerson, J., Bird, P. K., McEachan, R. R., Pickett, K. E., Waiblinger, D., Uphoff, E., . . . Sahota, P. (2016). Born in Bradford's Better Start: an experimental birth cohort study to evaluate the impact of early life interventions. *BMC Public Health*, *16*(1), 711.

Donnelly, J. E., Hillman, C. H., Castelli, D., Etnier, J. L., Lee, S., Tomporowski, P., . . . Szabo-Reed, A. N. (2016). Physical activity, fitness, cognitive function, and academic achievement in children: A systematic review. *Medicine and Science in Sports and Exercise*, *48*, 1197. https://dx.doi.org/10.1249%2FMSS.0000000000000901

Duncan, G. J. & National Institute of Child Health and Human Development Early Child Care Research Network (2003). Modeling the impacts of child care quality on children's preschool cognitive development. *Child Development*, *74*, 1454–1475. https://doi.org/10.1111/1467-8624.00617

Duncan, G. J., & Gibson-Davis, C. M. (2006). Connecting child care quality to child outcomes: Drawing policy lessons from nonexperimental data. *Evaluation Review*, *30*, 611–630. https://doi.org/10.1177/0193841X06291530

Economic and Social Research Council (2017). *Longitudinal studies strategic review: 2017 Report to the Economic and Social Research Council [PDF File]*. Retrieved from https://esrc.ukri.org/files/news-events-and-publications/publications/longitudinal-studies-strategic-review-2017/

Ferrie, J. E. (2012). The irresistible rise of the Cohort Profile. *International Journal of Epidemiology*, *41*, 899–904. https://doi.org/10.1093/ije/dys119

Hakim, C. (1982). Secondary analysis and the relationship between official and academic social research. *Sociology*, *16*(1), 12–28. https://doi.org/10.1177/0038038582016001005

Hutchison, D., & Schagen, I. P. (Eds.) (1994). *How reliable is National Curriculum assessment?*. Berkshire, UK: National Foundation for Education Research.

Information Commissioner's Office (2018). *Guide to General Data Project Regulation (GDPR) [PDF file]*. Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/

Jay, M. A., Mc Grath-Lone, L., & Gilbert, R. (2019). Data resource: The National Pupil Database (NPD). *International Journal of Population Data Science*, *4*(1), 1101. https://doi.org/10.23889/ijpds.v4i1.1101

Johnston, M. P. (2017). Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries*, *3*, 619–626.

Komarraju, M., Karau, S. J., Schmeck, R. R., & Avdic, A. (2011). The Big Five personality traits, learning styles, and academic achievement. *Personality and Individual Differences*, *51*, 472–477. https://doi.org/10.1016/j.paid.2011.04.019

Krapohl, E., & Plomin, R. (2016). Genetic link between family socioeconomic status and children's educational achievement estimated from genome-wide SNPs. *Molecular Psychiatry*, *21*, 437–443. https://doi.org/10.1038/mp.2015.2

Lupton, R., Burchardt, T., Fitzgerald, A., Hills, J., McKnight, A., Obolenskaya, P., . . . Vizard, P. (2015). *The Coalition's Social Policy Record: Policy, spending and outcomes 2010-2015*. Social Policy in a Cold Climate, Research Report 4, January 2015.

Matthews, J. S., Ponitz, C. C., & Morrison, F. J. (2009). Early gender differences in self-regulation and academic achievement. *Journal of Educational Psychology*, *101*, 689–704. https://doi.org/10.1037/a0014240

McCabe, A., Nic An Fhailí, S., O'Sullivan, R., Brenner, M., Gannon, B., Ryan, J., . . . Wakai, A. (2019). Development and validation of a data dictionary for a feasibility analysis of emergency department key performance indicators. *International Journal of Medical Informatics*, *126*, 59–64. https://doi.org/10.1016/j.ijmedinf.2019.01.015

Medical Research Council (2014). *Maximising the Value of UK Population Cohorts - Strategic Review of the Largest UK Population Cohort Studies [PDF file]*. Retrieved from https://mrc.ukri.org/publications/browse/maximising-the-value-of-uk-population-cohorts/

Nikulina, V., Widom, C. S., & Czaja, S. (2011). The role of childhood neglect and childhood poverty in predicting mental health, academic achievement and crime in adulthood. *American Journal of Community Psychology*, *48*(3–4), 309–321. https://doi.org/10.1007/s10464-010-9385-y

Northstone, K., Lewcock, M., Groom, A., Boyd, A., Macleod, J., Timpson, N., & Wells, N. (2019). The Avon Longitudinal Study of Parents and Children (ALSPAC): An update on the enrolled sample of index children in 2019. *Wellcome Open Research*, *4*, 51. https://dx.doi.org/10.12688%2Fwellcomeopenres.15132.1

Oliver, B. R., & Plomin, R. (2007). Twins' Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Research and Human Genetics*, *10*(1), 96–105. https://doi.org/10.1375/twin.10.1.96

Pell, J. (2014). *Maximising the Value of UK Population Cohorts: MRC Strategic Review of the Largest UK Population Cohort Studies*. Project Report. Medical Research Council.

Pell, J. P., Valentine, J., & Inskip, H. (2014). One in 30 people in the UK take part in cohort studies. *Lancet*, *383*, 1015–1016. https://doi.org/10.1016/S0140-6736(14)60412-8

Polderman, T. J. C., Boomsma, D. I., Bartels, M., Verhulst, F. C., & Huizink, A. C. (2010). A systematic review of prospective studies on attention problems and academic achievement. *Acta Psychiatrica Scandinavica*, *122*, 271–284. https://doi.org/10.1111/j.1600-0447.2010.01568.x

Queirós, A., Faria, D., & Almeida, F. (2017). Strengths and limitations of qualitative and quantitative research methods. *European Journal of Education Studies*, *3*, 369–387. https://doi.org/10.5281/zenodo.887089

Rimfeld, K., Malanchini, M., Spargo, T., Spickernell, G., Selzam, S., McMillan, A., … Plomin, R. (2019). Twins early development study: A genetically sensitive investigation into behavioral and cognitive development from infancy to emerging adulthood. *Twin Research and Human Genetics*, *22*, 508–513. https://doi.org/10.1017/thg.2019.56

Schoon, I., Jones, E., Cheng, H., & Maughan, B. (2012). Family hardship, family instability, and cognitive development. *Journal of Epidemiology and Community Health*, *66*, 716–722. https://doi.org/10.1136/jech.2010.121228

Siddiqui, N. (2019). Using secondary data in education research. *Social Research Update*, *68*, 1–4.

Smith, A. K., Ayanian, J. Z., Covinsky, K. E., Landon, B. E., McCarthy, E. P., Wee, C. C., & Steinman, M. A. (2011). Conducting high-value secondary dataset analysis: An introductory guide and resources. *Journal of General Internal Medicine*, *26*, 920–929. https://dx.doi.org/10.1007%2Fs11606-010-1621-5

Smith, E. (2008). *Using secondary data in educational and social research*. Maidenhead, UK: McGraw-Hill Education.

Taggart, B., Sylva, K., Melhuish, E., Sammons, P., & Siraj, I. (2015). *Effective pre-school, primary and secondary education project (EPPSE 3–16+): How pre-school influences children and young people's attainment and developmental outcomes over time [PDF file]*. Department for Education, Research Brief, January 2015. Retrieved from https://pdfs.semanticscholar.org/a184/228deaa434a7786c2516a19c47021d42fa60.pdf

Wellcome's Longitudinal Population Studies Working Group (2017). *Longitudinal Population Study Strategy: Wellcome's Longitudinal population Studies Working Group*. Wellcome Longitudinal Population Studies Strategy, July 2017. Retrieved from https://wellcome.ac.uk/sites/default/files/longitudinal-population-studies-strategy_0.pdf

Vaishnav, R. S., & Chirayu, K. C. (2013). Learning style and academic achievement of secondary school students. *Voice of Research*, *1*, 1–4.

von Stumm, S., Smith-Woolley, E., Ayorech, Z., McMillan, A., Rimfeld, K., Dale, P. S., & Plomin, R. (2020). Predicting educational achievement from genomic measures and socioeconomic status. *Developmental Science*, *23*, e12925. https://doi.org/10.1111/desc.12925

Watson, N., & Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In R. M. Groves, G. Kalton, J. N. K. Rao, N. Schwarz, C. Skinner, & P. Lynn (Eds.) *Methodology of longitudinal surveys*. https://doi.org/10.1002/9780470743874.ch10

Weis, M., Heikamp, T., & Trommsdorff, G. (2013). Gender differences in school achievement: The role of self-regulation. *Frontiers in Psychology*, *4*, 442. https://doi.org/10.3389/fpsyg.2013.00442

Wright, J., Small, N., Raynor, P., Tuffnell, D., Bhopal, R., Cameron, N., … West, J. (2013). Cohort profile: The Born in Bradford multi-ethnic family cohort study. *International Journal of Epidemiology*, *42*, 978–991. https://doi.org/10.1093/ije/dys112

Received 2 June 2020; revised version received 7 October 2020