IOE, Faculty of Education and Society

An investigation of the consistency of GCSE qualifications data in administrative educational records and a national social survey

CLS working paper number 2025/3

By Sarah Stopforth[1], Roxanne Connelly[2], Vernon Gayle[3]

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

[1] University of York

[2] University of Edinburgh

[3] University of Edinburgh

CENTRE FOR LONGITUDINAL STUDIES

Economic and Social Research Council

Contact the author

Vernon Gayle

University of Edinburgh

vernon.gayle@ed.ac.uk

This working paper was first published in July 2025 by the UCL Centre for Longitudinal Studies.

UCL Social Research Institute

University College London

20 Bedford Way

London WC1H 0AL

www.cls.ucl.ac.uk

The UCL Centre for Longitudinal Studies (CLS) is an Economic and Social Research Council (ESRC) Resource Centre based at the UCL Social Research Institute, University College London. It is home to a unique series of UK national cohort studies.
For more information, visit www.cls.ucl.ac.uk.

This document is available in alternative formats. Please contact the Centre for Longitudinal Studies:
Email: clsdata@ucl.ac.uk

Disclaimer

This working paper has not been subject to peer review.

CLS working papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Any opinions expressed here are those of the author(s) and not those of the UCL Centre for Longitudinal Studies (CLS), the UCL Social Research Institute, University College London, or the Economic and Social Research Council.

How to cite this paper

Stopforth, S., Connelly, R., Gayle, V. (2025) *An investigation of the consistency of GCSE qualifications data in administrative educational records and a national social survey* CLS Working Paper 2025/3. London: UCL Centre for Longitudinal Studies.

Abstract

This working paper reports on General Certificate of Secondary Education (GCSE) data within the Millennium Cohort Study (MCS). Data on school GCSE results were collected from MCS cohort members via a self-report instrument in sweep 7 of the survey (age 17). For a subset of respondents (with parental consent) administrative educational records in the National Pupil Database (NPD) containing GCSE results have been linked to the main MCS survey. Valid and reliable educational qualifications data are essential for a wide array of research fields.

The focus of this working paper is an investigation of the consistency of these two sources of GCSE data. Statistical models are used to formally examine consistency between the GCSE results reported in the survey and the administrative data, and we conduct a concise sensitivity analysis of alternative GCSE attainment measures. We observed marked inconsistencies between self-reported GCSE results in the survey and the GCSE results recorded in the administrative data. We conclude that in the analysis of GCSE attainment, the use of either source of GCSE data would not have resulted in vastly different substantive conclusions. We warn researchers that there is no guarantee that the inconsistencies in GCSE reporting will be negligible in all analyses. Our central recommendation is that researchers should use the administrative GCSE records whenever it is practicable.

The key messages for researchers who intend to use GCSE data in the MCS are:

The GCSE data within the MCS resources offer unparalleled resources for studying education in contemporary Britain.

The self-reported GCSE data provide a useful source of information. However, our investigations suggest that these data are not ‘research ready’ and require some processing and checking.

We advise researchers to access the linked GCSE data from the National Pupil Database (NPD). Our investigations indicate that the NPD data are also not immediately ‘research ready’ and require some processing and checking.

Extended exploration of the data indicates that some of the deposited NPD measures are sub-optimally processed (i.e. they are not 'research ready'), and therefore unsuitable for immediate use in analyses (e.g. in the pupil-level dataset).

We recommend that data analysts use the raw NPD data (i.e. the subject and qualification codes) to derive valid and reliable measures that are suitable for the specific analysis.

There have been a number of unforeseen data management challenges in the preparation of both the self-reported data and administrative records. In order to improve the usability of these data for future researchers, we have made our statistical code available here: https://osf.io/gs6m2/. Our code demonstrates the entire data wrangling process for the linked administrative data, the self-reported GCSE qualifications data, and the analysis undertaken in this paper.

In future data collections of educational qualifications, we recommend the use of a more structured approach to the collection of detailed educational measures. When self-report data collection instruments are employed, we recommend developing a qualifications grid or establishing an external validation protocol based on official examination transcripts.

Acknowledgement

We thank the participants of the Millennium Cohort Study. We are also grateful to the Centre for Longitudinal Studies (CLS), UCL Social Research Institute, for the use of these data and to the UK Data Service for making them available. However, neither CLS nor the UK Data Service bear any responsibility for the analysis or interpretation of these data. We would like to thank Dr Chris Playford for helpful comments on an earlier version of this paper, and we would also like to thank Professor Emla Fitzsimons and Dr Richard Silverwood for helpful comments on this working paper.

This work was supported by the Economic and Social Research Council, grant number: ES/X012085/1.

Introduction

The UK has an unparalleled collection of nationally representative birth cohort studies that track individuals across the life course (Wadsworth and Bynner, 2011). The Millennium Cohort Study (MCS) is an exceptional resource for studying young people growing up in the 21st Century. The MCS is an important multidisciplinary data resource which captures information on early family context, child development, outcomes in childhood and later in adolescence, and subsequently on outcomes in adulthood (Hansen, 2014). The MCS maintains the essential design of previous British birth cohort studies but has a range of new features including a sample of births from across the year, a special sample of deprived areas, and a large sample that supports analyses at UK territorial levels (Smith and Joshi, 2002).

The MCS is a study of children born in 2000-2002 (Connelly and Platt, 2014). Members of the MCS are the first group within a large-scale nationally representative birth cohort to study for GCSE qualifications. An exciting innovation that sets the MCS apart from earlier British birth cohort studies, is that individual level pupil information from the National Pupil Database (NPD) has been linked to the survey data. NPD data combined with the MCS survey data provides an incomparably powerful resource for studying school qualifications and educational inequalities in England.

Data on school GCSE results were collected from the young people via a self-reporting instrument in MCS sweep 7 (around age 17) (University of London et al., 2023). General Certificates of Secondary Education (GCSEs) are the main school qualifications that are undertaken by most school pupils in England, usually when they are aged 15-16 (Gill, 2016). GCSE data is inherently complex because English pupils undertake a selection of GCSE qualifications across a range of individual subjects, and each GCSE subject is awarded a separate grade. There are a small number of GCSE subjects which can be awarded as a double award, which are worth two GCSEs and will result in two adjacent GCSE grades. Pupils undertake a mixture of compulsory and optional GCSE subjects, and their portfolio of results is highly individualised (Connelly et al., 2016a).

Self-reported data is ubiquitous in the social sciences, but it is far from routine for social surveys to collect the level of detail (i.e. subject-specific grades) as collected in sweep 7 of the MCS survey. Increasingly, the granularity of individual-level qualifications data requires the use of administrative educational records. The official nature of administrative records may lead to the false presumption that they are more accurate than self-report data (Adriaans et al., 2020). Whilst advising researchers to think carefully about the quality of administrative data for each specific research question, Goerge and Lee (2001) recommend that researchers assess data quality by comparing the administrative data with an alternative source. It is not often the case, however, that researchers have access to another data source which can provide a suitable comparison.

In this paper, we are in the methodologically fortunate position to consider the consistency of GCSE qualifications which are self-reported by young people in the Millennium Cohort Study (MCS) (University of London et al., 2023) with GCSE qualifications recorded in official administrative educational records in the National Pupil Database (NPD) (University College London et al., 2021). Pupils sit their GCSE examinations at the age of 15 or 16. The MCS asked cohort members to self-report their qualifications in sweep 7 of the study, when they were aged 17. The linked data therefore provides an innovative opportunity to test the consistency of the official administrative records with GCSE qualifications which were self-reported in the subsequent year.

Our analysis will address three key questions:

1. Are self-reported qualifications in the MCS consistent with official administrative educational records in the NPD?
2. Are cohort member characteristics associated with patterns of inconsistency between self-reported and administrative records?
3. What potential impact do data disagreements have on empirical analyses?

We offer some practicable methodological advice for social researchers analysing the MCS self-reported GCSE data, and administrative data in the NPD.

Consistency of administrative educational data and social survey records

Administrative data are commonly collected for the three key purposes of registration, transaction, and record keeping (Woollard, 2014). A promising development in the UK data infrastructure has been the commitment from agencies to enable better access to administrative data for social research (Administrative Data Taskforce, 2012). The potential benefits of administrative datasets have been well rehearsed (see Card et al., 2010). Although not collected for research purposes, administrative data provide rich detail at the individual-level which can be successfully exploited for research (Jones et al., 2018). A frequently overlooked aspect of undertaking social research with administrative data is that these resources contain fewer explanatory variables than the wide array that are common in social surveys (Connelly et al., 2016b).

The Department for Education collects official administrative GCSE qualifications data for young people in England in the National Pupil Database (NPD) (for an outline see Jay et al., 2019). A fundamental limitation of the NPD for undertaking sociological research is that it is a set of purely administrative records and does not contain the vector of explanatory variables that are routinely collected in social surveys. This is a common limitation of administrative datasets (Playford et al., 2016). A specific challenge for studying social inequalities in education is that the NPD does not include sociological measures such as parental social class and parental education.

In the UK, there has been a promising advent of linked administrative records to large-scale, nationally representative social surveys which facilitate rich analyses at the individual-level. However, the nature of administrative data means that there are strict access arrangements, which is often a lengthy process. Once approval is granted, researchers must undergo specific training, undertake their research in secure settings with restricted network and internet access, and adhere to strict disclosure control procedures for any outputs they wish to remove from the secure setting; any breaches of data access rules may be punishable via sanctions or legal

action (Harron et al., 2017). This adds complexity to the research process, and can cause issues for the data analysis workflow.

Social surveys often collect information about educational qualifications. In longitudinal (i.e., repeated contacts) studies, information about a person's educational history is usually collected at the initial interview, and respondents are asked whether they have any additional qualifications at each subsequent wave. Often, the level of detail relating to the qualifications does not go beyond asking the respondent the number of qualifications they have obtained at a certain level, for example, the number of GCSEs. The qualifications information collected in the main survey of the MCS is much more detailed, and subject-specific grades are available with the main survey dataset under an End User Licence. The availability of specific grade information within a mainstream social survey broadens access to these important social science variables as both explanatory and outcome variables for a wide range of research questions. A number of papers are emerging which make use of the MCS self-reported GCSE and iGCSE data (for example, Anders et al., 2024, Elliot Major and Parsons, 2022, Walker and Gamble, 2023). A useful feature of the self-reported data is that it enables the examination of educational attainment for both private school and state school pupils (see Anders et al., 2024).

There are no comprehensive studies which specifically compare the administrative records to self-reported educational data in the UK. The relationship between administrative educational records and self-reported data has been studied in other countries. Kuncel et al. (2005) undertook a meta-analysis of research examining the accuracy of self-reported grade point average data in the United States and found that correlations between self-reported and 'actual' grades were higher for those with higher 'actual' grades. Sticca et al. (2017) examined the consistency of the European 6-point grading scale and found high reliability of self-reported grades, with some differences by school subject. Self-reporting of Mathematics was more consistent than languages, but the absolute levels of over-reporting and under-reporting was low. Adriaans et al. (2020) analysed information on school leaving certificates and vocational and university degrees using linked social security and survey data in Germany and found large inconsistencies between the data sources. Missingness

and inconsistencies between administrative and self-reported qualifications were related to the type of degree obtained.

Goerge and Lee (2001) noted that information which is not central to the administrative process, and the purpose for its collection, may not be of high quality. For that reason, we must not presume, *a priori*, that administrative records are accurate. In this instance, the collection of examination results is the central administrative process, and therefore we believe we can be fairly confident in the data quality of the administrative records held in the National Pupil Database. However, challenges exist in linked administrative data with regards linkage errors, for example if records cannot be linked or are linked incorrectly (Harron et al., 2017).

Young people receive an official transcript from each exam board detailing each GCSE subject qualification and the grades they have achieved. As young people venture into further education and employment, they will be required to self-report the qualifications they have acquired as part of their educational histories in their university and job applications. In this paper, we directly compare the self-reported responses with the official records held in the NPD. The findings of this paper will have important implications for both data collectors and researchers when using self-reported qualifications data.

Data and Methods

Linked administrative and social survey data

We analyse data from the UK Millennium Cohort Study (MCS) (Connelly and Platt, 2014). The MCS is a nationally representative, large-scale, birth cohort study of almost 19,000 children born in the UK between September 2000 and January 2002. English cohort members in the MCS were born between 1st September 2000 and 31st August 2001, and constitute a single academic school year cohort. The cohort members entered Reception class in school year 2005/2006, and sat their GCSE qualifications in school year 2016/2017.

In MCS sweep 4 (when cohort members were aged 7), parents or carers were invited to consent to administrative data from the National Pupil Database (NPD) being linked to the MCS survey data (Rihal and Gomes, 2021). In 2018, NPD

records were linked to all consenting MCS cohort members who were resident in England, providing access to data from Key Stage 1, Key Stage 2, and Key Stage 4. The successful linkage rate for consenting cohort members was 99.4% (a total of 8438 out of 8489) (Rihal and Gomes, 2021).

In MCS sweep 7 (calendar year 2018/2019), respondents self-reported their educational qualifications, including their recently awarded GCSE results. To assess the consistency of GCSE qualifications data, we use the MCS linked with the NPD (University College London et al., 2021) for the administrative records, and sweep 7 of the main MCS survey for the self-reported qualifications (University of London et al., 2023). The MCS main survey is accessible from the UK Data Service (UKDS) under a standard End User Licence. The linked administrative data are deemed too confidential and sensitive to be made available under a standard End User License (i.e. they are designated as controlled data). These data must be accessed via the UKDS Secure Lab and results undergo a statistical disclosure control process before they are released. We have made our full research code available in order to enable researchers to duplicate our results and to replicate the measures for other research enterprises. The code is available here: https://osf.io/gs6m2/.

Sample

The analysis is undertaken on those who have matched records (i.e. they have self-reported their qualifications in the MCS and they also have linked administrative records in the NPD). Our sample is defined as MCS members living in England at sweep 7 who have self-reported their GCSEs, have linked NPD records, and have at least 1 GCSE grade recorded in both datasets. Table 1 presents the available sample sizes in both datasets. The corresponding sample with at least 1 GCSE grade recorded in both datasets is 5,410 cohort members.[1] Those who have reported their qualifications in the MCS but do not have an NPD record are

[1] There are 36 people who have records in both the MCS and NPD (and have therefore been successfully linked), but they do not have any GCSE grades recorded in one of the datasets. We cannot disaggregate these 36 people due to statistical disclosure control. The majority of the 36 are people in the MCS who report that they have GCSEs, but report that they 'don't know' their grade for every subject they mention. There are also a very small number of people in the NPD who have been successfully linked, but there is no record of them having taken GCSEs.

potentially unsuccessful data linkages, or perhaps did not give consent for their data to be linked. Those who have an NPD record but did not report their qualifications in the MCS potentially did not answer in sweep 7, refused to answer the questions, or perhaps did not sit GCSEs.

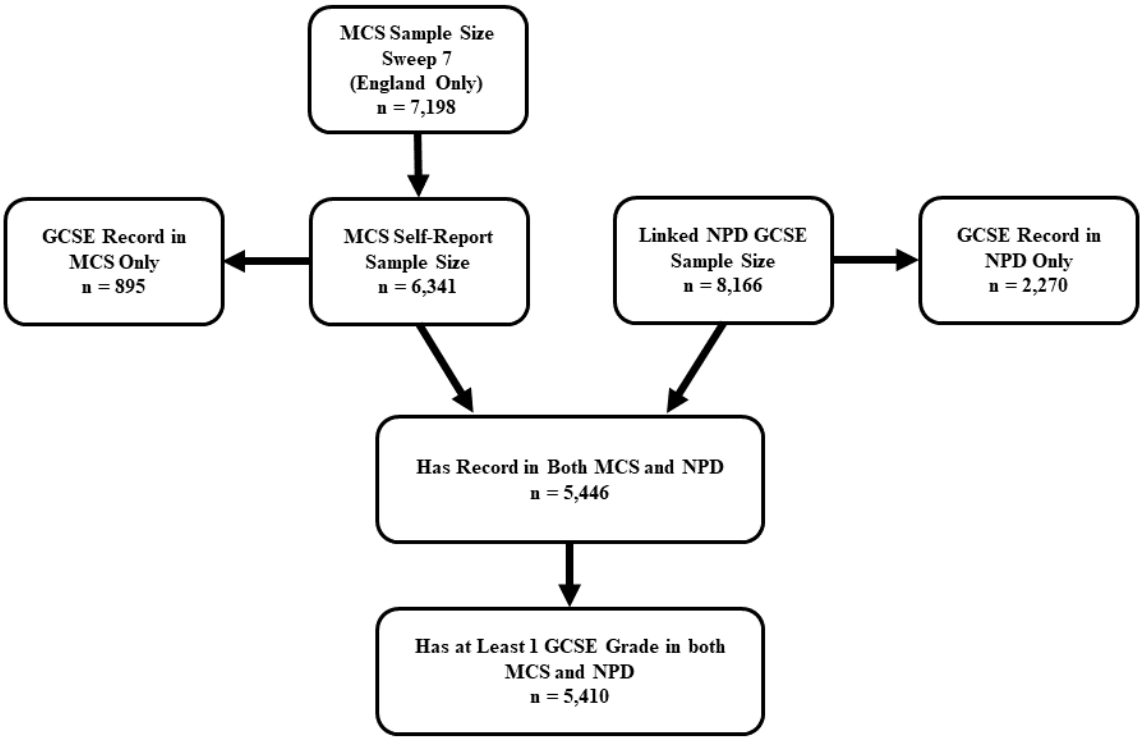


Figure 1: Sample sizes available in the self-reported dataset and administrative records

GCSE measures

From the introduction of GCSEs in the late 1980s until the mid-2010s, individual GCSE subjects were given an alphabetical grade. The highest grade was an A, and the lowest grade a G. From 1994, a higher grade of A* was introduced (Yang and Woodhouse, 2001). Following reforms in 2016, the GCSE grading system has changed in England. A numeric grading system was phased in over several years, starting in academic year 2016/2017. Grade 9 is the new highest grade and grade 1 the lowest (Ofqual, 2018). Summer 2017 was the first examination season of the phased numerical grading system, meaning that a typical pupil was awarded numerical grades in English Language, English Literature, and Mathematics, and alphabetical grades in all other subjects. The majority of the English Millennium

Cohort Study members sat their GCSEs in the summer of 2017, when two different grading systems were simultaneously in operation. This provides a specific additional complexity to summarising GCSE grades for the cohort members in this study.

In the datasets, we have two indicators of the 'true' GCSE result, one which is self-reported, and one which is administrative. By comparing the consistency of these scores between the two datasets, we can ascertain the validity and reliability of these measures, and the impact of any inconsistencies on the conclusions drawn from empirical analyses. There are several ways to measure overall GCSE outcomes. No single measure has been universally agreed upon or routinely adopted for social research. We operationalise a range of plausible measures which summarise overall GCSE outcomes (see Table 1).

Table 1: GCSE summary measures

Measure	GCSE measure	Description
1	Number of GCSEs	Total count of GCSE qualifications recorded
2	Number of 'good' passes	Total count of GCSE qualifications recorded at grades A*-C or 9-4
3	5+ 'good' passes	Binary measure of whether achieved 5 or more GCSE qualifications at grades A*-C or 9-4
4	5+ 'good' passes including English and Mathematics	Binary measure of whether achieved 5 or more GCSE qualifications at grades A*-C or 9-4, including in English Language or English Literature, and Mathematics
5	Combined score	Total GCSE points score based on a combined score of the numeric and alphabetical grading systems
6	Combined score (capped)	Combined score capped at the equivalent of the best 8 subjects (an overall cap of 67 points, i.e. 27 numeric points in English Language, English Literature, and Mathematics, plus 40 alphabetical points in 5 additional subjects)
7	Interpolated score	Total GCSE points score based on an interpolated score of the numeric and alphabetical grading systems (see Table 2)
8	Interpolated score (capped)	Interpolated score capped at the equivalent of the best 8 subjects (an overall cap of 65.5 points, i.e. 25.5 [8.5*3] numeric points in English Language, English Literature, and Mathematics, plus 40 [8*5] alphabetical points in 5 additional subjects)
9	Mean score	Mean GCSE points score based on the interpolated score

Measure 1 is the total number of GCSEs. This is the most rudimentary measure of consistency between the two datasets. We derive this measure by counting the number of subject grades recorded in each dataset.

Measure 2 is the number of ‘good’ passes. ‘Good’ passes can be understood as achieving an alphabetical grade of A*-C or a numeric grade of 9-4. Although a numeric grade of 9-5 is considered a ‘strong’ pass and 9-4 is considered a ‘standard’ pass, Ofqual (2018) advised that the bottom of a grade 4 was designed to be the equivalent of the bottom of an old grade C. Therefore, we use the benchmark of a grade 4 to aid comparability across grading systems.

Measure 3 is a binary measure of whether the cohort member achieved 5 or more ‘good’ passes or not. The achievement of 5 or more A*-Cs was often used as a benchmark measure in policy research and government statistics (Leckie and Goldstein, 2009).

Measure 4 is a binary measure of whether the cohort member achieved 5 or more ‘good’ passes including English and Mathematics or not. More recently, government league tables have included a measure of the percentage of pupils gaining 5 or more GCSEs at grades A*- C including English and Mathematics (Leckie and Goldstein, 2017).

Measure 5 is a combined score of numeric and alphabetical point scores. For this academic cohort, two grading systems were in operation. We derived separate scores for numeric and alphabetical grades. The numeric scoring system preserves the discrete, ordered categories on a numeric scale, whereby each grade 9 receives 9 points, each grade 8 receives 8 points and so on. The numeric scale ranges from 0 to a maximum of 27 points. We have capped the maximum numeric points score at 27 points to reflect the new numeric grades for English Language, English Literature, and Mathematics. We convert each alphabetical grade into a numeric score, following the established guidance provided in Yang and Woodhouse (2001), whereby each A* receives 8 points, each A receives 7 points, and so on. The two scoring schemes are not identical, however this transformation maximises functional equivalence.

Measure 6 is a capped measure of the combined score at the equivalent of the best 8 subjects (for this cohort, we consider this to be 27 numeric points in English Language, English Literature, and Mathematics, plus 40 alphabetical points in 5 additional subjects, with an overall cap of 67 points).

Measure 7 is a total points score using an alternative scoring system which interpolates the numeric and alphabetical scoring systems into one scale. We begin with the standard grade conversion used for alphabetical grades as described in Yang and Woodhouse (2001). We then interpolate the numeric grades into this system. Following Ofqual (2018) advice, the bottom of a grade 7 is equivalent to the bottom of an A grade, the bottom of a grade 4 is equivalent to the bottom of a C grade, and the bottom of a grade 1 is equivalent to a G grade. This is reflected in the same interpolated score being given to these grades. Table 2 presents the interpolated scale. We acknowledge that it is not possible for a pupil to achieve higher than a score of 8 for any subject using the alphabetical grading system, whereas the highest score is an 8.5 for subjects with the numeric grading system. The new grade 9 is deliberately 'higher' than the old A* grade (Ofqual, 2018), but this is a function of the grading system itself and not a reflection of the pupil's achievement.

Measure 8 is a capped points score measure at the equivalent of the best 8 subjects (for this cohort, we consider this to be 25.5 points for numeric grades in English Language, English Literature, and Mathematics [8.5*3], plus 40 alphabetical points in 5 additional subjects [8*5], with an overall cap of 65.5 points).

Measure 9 is an individual pupil's mean score based on the full interpolated scale (i.e. measure 7 divided by measure 1).

Table 2: Interpolated scoring system for numeric (new) and alphabetical (old) grading systems

New	Interpolated	Old
9	8.5	
	8	A*
8	7.5	
7	7	A
6	6.5	
	6	B
5	5.5	
4	5	C
	4	D
3	3.5	
	3	E
2	2.5	
	2	F
1	1	G

Analytical strategy

In order to examine the consistency of the self-reported data with the official administrative educational records, we compare each alternative summary GCSE measure outlined above. We report how many cases correspond exactly, and the correlation of measures in the self-reported and administrative records. Next, we attempt to understand the differences by concentrating on the reasons for inconsistency of reporting grades for the three core subjects of English Language, English Literature, and Mathematics.

The association between cohort member characteristics and differences between self-reported and administrative records are investigated using a series of statistical models to evaluate patterns of inconsistency. Previous studies comparing administrative and self-reported educational data in other national contexts found that those with lower grades tended to report their grades less reliably (for example, Adriaans et al., 2020, Kuncel et al., 2005). We test whether inconsistencies are influenced by the grade achieved (as recorded in the administrative dataset). We also test whether sex and social class affect patterns of inconsistency between self-reported and administrative GCSE data (see Jerrim et al., 2019).

To examine the potential impact that the inconsistencies have on empirical analyses, we compare the results from models which use the summary GCSE measures as reported in the MCS and those recorded in the NPD. We include the additional explanatory variables sex, parental education level, and parental social class (measured by the National Statistics Socio-Economic Classification, NS-SEC (Rose and Pevalin, 2003)). We appropriately adjust the analyses for complex survey design and non-response in the MCS.

We derive explanatory variables from the main survey of the Millennium Cohort Study. We derive parental social class and parental education information from sweep 6 (age 14) and sex from sweep 7 (age 17). Table 3 presents the descriptive statistics. Using social survey data inevitably involves a degree of missing data (Carpenter and Kenward, 2013). There are missing data on the survey variables parental NS-SEC and parental education, and there are 970 cases with incomplete records. Our complete records analytical sample for this phase of the analysis is n=4,440.

Table 3: Descriptive statistics for explanatory variables

	Frequency	Percentage	Adjusted Percentage+
Parental NS-SEC			
1.1 Large Employers and Higher Managerial	250	5.63	6.34
1.2 Higher Professional	659	14.84	16.54
2 Lower managerial and professional	1,263	28.45	29.79
3 Intermediate	507	11.42	11.36
4 Small employers and own account	727	16.37	15.81
5 Lower Supervisory and Technical	276	6.22	5.55
6 Semi-Routine	453	10.20	8.96
7 Routine	305	6.87	5.65
Parental Education			
No degree	3,491	78.63	78.41
Degree	949	21.37	21.59
Sex			
Male	2,143	48.27	48.58
Female	2,297	51.73	51.42
Total	4,440	100.00	100.00

Note: + Adjusted for the complex sample design and non-response of the Millennium Cohort Study.

Data management challenges

Administrative data that are not collected for research purposes typically require a large amount of cleaning before they are ready for analyses (Connelly et al., 2016c). The general nature of large-scale surveys means that even the best curated datasets will require some data wrangling prior to specific data analysis. As researchers unconnected with the original data collection exercises, we found that the self-reported qualifications data in sweep 7 of the Millennium Cohort Study were in a format that does not support immediate data analysis. Despite great effort and consideration of alternative modes of data collection, we reflect that the approach taken has led to some unintended consequences. There were unforeseen data management challenges in the preparation of both the self-reported data and administrative records. In order to improve the usability of these data for future researchers, we draw upon research transparency and reproducibility guidance (for example, see Gayle and Connelly, 2022). We deposit our statistical code alongside this paper, which outlines the entire data wrangling process for the linked

administrative data, the self-reported GCSE qualifications data, and the analysis undertaken in this paper which compares the two. In this code, we have taken the prudent step of adding data signatures to each dataset used in this project (see Gould, 2006). This practice ensures that the original research team and other data users are working with identical datasets.

In the self-reported qualifications interview in sweep 7 of the main MCS survey, cohort members were asked 'Which, if any, of the qualifications on this card do you have? Please include all qualifications regardless of grade'. Interviewers were told to only include qualifications that the young person had results for (and therefore were not currently waiting for results), and to include the original grade if the young person was studying for a resit. Short course versions of the qualification were reported as an additional qualification, and if a qualification was a double award, it was flagged in the dataset. GCSE was the first option on the qualification list. Respondents were then asked how many GCSEs they had, regardless of the grade. If the number of GCSEs was over a certain threshold (which is not disclosed in the documentation), the interviewer was asked to correct it. Respondents were then asked to 'tell us the subject(s) you studied and the grade(s) you got for your GCSE(s)', with an interviewer prompt that the respondent should choose the most similar title, if the exact course title did not appear in the list. The list of available subjects to choose from was therefore pre-populated. Two soft checks were performed: one to make sure the number of GCSEs reported corresponded with the number of GCSEs they said they had with options to 'amend' or 'continue', e.g. if one or more subjects were double awards. The second soft check occurred if the respondent selected the same subject twice, and asked the interviewer to either 'amend' or 'continue'. Interviewers did not undertake external validity checks of the cohort members' answers, for example by checking their official examination transcripts.

Despite the interviewer soft check of subjects being mentioned twice, many subjects were still mentioned more than once by respondents. After seeking advice from the data support team, the reasons for duplicate mentions of subjects in the Millennium Cohort Study were not clear. The data documentation is not detailed enough to explain why some subjects appear multiple times for the same individual. Potential explanations proffered by the data team included the use of aggregated subject

categories meaning that several GCSE subjects were combined under the same overall subject option in the drop-down menu, interviewer error, or young people reporting modular instead of overall results. However, there is no way to verify these explanations in the deposited data.

A similar peculiarity we faced was the surprising volume of subjects flagged as ‘double awards’. Typically, a double award would count for two separate GCSEs. We suspect there are errors in the flagging of double awards, largely because people have self-reported double awards in subjects which do not ordinarily have this provision.

An additional difficulty with using the self-reported data is the structure of the deposited dataset. Cohort members reported their GCSE qualifications in the order in which they were able to recall their results. This means that there was not a standardised order to the subject list. In the dataset, each person not only has an individualised GCSE subject portfolio, but also an individualised order of subjects that they could recall. A hypothetical example to illustrate this is shown in Table 4. In this hypothetical example, Person 1 reported Mathematics first, Physics second, and Chemistry third. Person 2 also reported Chemistry, but this was the second subject they recalled. Person 2 reported English Literature third, whereas Person 3 reported English Literature first. A technical problem with this data collection approach meant that the structure of the raw data required a lot of initial data management to make it useable. Young people typically sit 9 GCSE examinations and have separate subject-specific GCSE grades, within which they are likely to have sat modular examinations, or a set of papers contributing to an overall GCSE subject grade. Relying on young people to recall their GCSE qualifications without any prompts could plausibly lead to forgetting to mention some subjects, or misreporting modular instead of overall GCSE grade results only.

Table 4: Hypothetical data structure for reporting GCSE subjects

Person	Mention	Subject
1	1	Mathematics
1	2	Physics
1	3	Chemistry
2	1	Sociology
2	2	Chemistry
2	3	English Literature
3	1	English Literature
3	2	Psychology
3	3	Biology

The NPD extract that is made available with the MCS contains a mixture of individual level raw GCSE data and pupil level summary statistics. A cursory examination of the summary statistics revealed two immediate problems. First, there was a large amount of missing data for core subjects, for example the variable for the full GCSE grade achieved in English Literature contains alphabetical grades only and is therefore mostly comprised of missing data, discounting the majority of young people in this academic cohort with numerical grades in English Literature. Second, there measures of the full GCSE grade for all GCSE subjects. For example, the full GCSE grade achieved in the core compulsory subject of English Language was not deposited in the linked pupil level dataset. We also found that some subjects which had been self-reported in the MCS did not have an equivalent variable in the deposited NPD extract, e.g. Sociology. Therefore, the data deposited in the GCSE pupil level dataset is not immediately ready for analyses. There is no documentation describing how the derived measures (summarising patterns of attainment) deposited in the dataset were constructed. We caution researchers from uncritically proceeding with the analysis of these data.

On querying this issue with the data team, we were advised to derive each GCSE subject grade variable using the unprocessed variables (subject codes, qualification

codes, and descriptions). We could then be certain that every GCSE subject recorded was represented in the analytical dataset for this cohort of young people.

The next irregularity we faced was the recording of the same subject multiple times in the administrative dataset. The vast majority of cases were due to the young person resitting the examination in a particular subject. There is no flag in the NPD for whether a result is the first grade achieved, or if the subject has been re-sat. Therefore, if taken at face value, many young people would appear to have two different grades in the same subject, and subsequently two separate GCSE qualifications in the same subject. It is possible to identify resits using information on the exam season and exam year that the student took the exam. Some duplicate mentions appeared to be administrative errors, for example the records contained the same information on most fields. We took the principled approach of only retaining the highest grade recorded for each duplicate subject.

In this paper, we have chosen to report the comparison between the Millennium Cohort Study and the National Pupil Database using all information as reported. It was not possible to verify reasons for duplicate subjects in the National Pupil Database, and therefore all duplicate mentions of a subject in the administrative data were double checked, and the highest grade was retained. However, without adequate documentation or additional information, we were unable to verify the reasons for subjects being mentioned more than once in the self-reported data. To investigate the impact of this decision, we first took the self-reported data at face value, i.e. retained all double awards and all duplicate mentions. These are the results reported below, as they are the most ‘faithful’ to the results reported by the MCS cohort members.

Next, we created four further datasets to check the robustness of results: (i) retain all double awards but remove all duplicates, (ii) remove all double awards but retain all duplicates, (iii) constrain double awards to only those available in the NPD and remove all duplicates, and (iv) remove all double awards and duplicates. It is worth noting that the number of perfect matches and correlations between the self-reported and administrative records are marginally, yet systematically, higher for all measures, for datasets where double awards are either ignored or constrained to only those available in the NPD. Unfortunately, due to very minor changes for some

of the measures, we are unable to release these additional analyses from the secure environment due to statistical disclosure control. However, we do provide our full research code as an accompaniment to this paper. Using all or no duplicate subject mentions have minor consequences for the number of perfect matches and correlations. Although many subjects have duplicate mentions, the frequency per subject is very low. The results of all three analyses (identifying inconsistencies, exploring patterns of inconsistencies, and the effects on substantive analyses) were unchanged in all derivations of the dataset.

Results

The consistency of GCSE results

In order to examine the consistency of the self-reported data with the official administrative educational records, we compare each alternative summary GCSE measure outlined above. We report how many cases correspond exactly, and the correlation of measures in the self-reported and administrative records. The direct matches between the survey records and the administrative records for each of the nine GCSE summary measures are reported in Table 5. The aggregate measures, 5 or more ‘good’ passes and 5 or more ‘good’ passes including English and Mathematics (measure 3 and measure 4), have the highest percentages of direct matches. These aggregate measures are broad indicators of GCSE attainment. The survey records and the administrative records directly match because of the low level of resolution in these measures. The summary measures with higher levels of resolution (measures 5 – 9) have a lower percentage of direct matches, however the correlation between the survey record and the administrative record are relatively strong.

Table 5: Comparison of GCSE summary measures in the self-reported and administrative datasets

GCSE measure	Perfect match		Correlation/ association
	Frequency	**Percentage**	
Number of GCSEs	2,175	40.20	r=0.58
Number of 'good' passes	2,711	50.11	r=0.90
5+ 'good' passes	5,052	93.38	V=0.83
5+ 'good' passes including English and Mathematics	5,041	93.18	V=0.85
Combined score	1,517	28.04	r=0.89
Combined score (capped)	2,064	38.15	r=0.87
Interpolated score	1,436	26.54	r=0.86
Interpolated score (capped)	1,925	35.58	r=0.86
Mean score	1,452	26.84	r=0.91
Total	5,410	100	

Note: r=Pearson's r correlation; V=Cramer's V

To understand why these inconsistencies occur, we focus on the different reasons for inconsistencies at the subject-level. As pupils sit highly individualised portfolios of GCSE subjects, it is not possible to know if a subject was not reported because the individual forgot to mention it, or because they did not sit the exam at all. This makes it difficult to understand how self-reported results differ from those held in administrative records. However, we can start to understand potential patterns by focusing on the core subjects of English Language, English Literature, and Mathematics. These subjects were compulsory for most, if not all, young people, and theoretically, these core subjects should have been mentioned by every respondent in the MCS, and should be recorded for every individual in the NPD. Table 6 presents the comparison of the three core subjects between the self-reported and administrative datasets, and patterns of inconsistencies.

Table 6: Comparison of reporting English Language, English Literature, and Mathematics GCSEs

Differences between self-reported and administrative records	English Language		English Literature		Mathematics	
	Frequency	Percentage	Frequency	Percentage	Frequency	Percentage
Perfect grade match	3,609	66.71	3,639	67.26	3,820	70.61
Missing grade in both MCS and NPD	119	2.20	173	3.20	141	2.61
Missing in MCS but recorded in NPD	468	8.65	475	8.78	381	7.04
Missing in NPD but reported in MCS	92	1.70	95	1.76	92	1.70
Reported one grade higher	273	5.05	247	4.57	217	4.01
Reported one grade lower	97	1.79	105	1.94	25	0.46
Converted numeric grade to correct alphabetical grade	412	7.62	382	7.06	467	8.63
Converted numeric to alphabetical but one grade higher	27	0.50	21	0.39	16	0.30
Converted numeric to alphabetical but one grade lower	155	2.87	125	2.31	124	2.29
More than one grade different	158	2.92	148	2.74	127	2.35
Total	5,410	100%	5,410	100%	5,410	100%

The patterns of inconsistencies between the two data sources are very similar for each of the three core subjects. Over two-thirds of young people who have records for English Language, English Literature, and Mathematics correctly report the result held in the administrative record. Between 7 and 9% of young people have an administrative record for these three core subjects, but do not self-report this information. This is surprising, given that everyone in this sample self-reports at least one other GCSE grade. This may suggest that respondents have simply forgotten to mention these three core subjects. Less than 2% of young people self-report a grade in these core subjects but there is no corresponding administrative record in the NPD.

A problem that is unique to this academic cohort is the phasing in of the new numeric grading system alongside the existing alphabetical grading system. Between 7 and 9% converted their numeric grade into an equivalent alphabetical grade. This was the first academic cohort to receive GCSEs using the new numeric grading system for these core subjects, and the rest of their GCSE results will have been on the alphabetical grading system. Between 4 and 5% misreported one grade higher in each of the three core subjects, whereas less than 2% misreported one grade lower. A minority of young people mis-converted their numeric grade to an alphabetical grade, with young people tending to under-estimate their grade by converting it to a lower alphabetical equivalent.

There are clear differences in the self-reporting of GCSE results compared with the administrative records held in the NPD. For this academic cohort, we would expect young people to have numeric grades in English Language, English Literature, and Mathematics, as well as alphabetical grades in 5-7 additional subjects, including at least one Science. This would be the typical and expected diet of GCSE results for young people sitting their examinations in summer 2017. When we compare this expectation in the two datasets (Table 7), 3,697 young people (68.3%) in the administrative data have this typical diet. However, only 2,189 of the same young people (40.5%) have this typical diet in the self-reported data.

Table 7: Achieved typical and expected results based on self-reported (MCS) and administrative (NPD) records

	Typical diet in NPD		
Typical diet in MCS	Yes	No	Total
Yes	2,031	158	2,189
No	1,666	1,555	3,221
Total	3,697	1,713	5,410

Note: Typical and expected results for this academic cohort are 8-10 GCSEs overall, including numeric grades in English Language, English Literature, and Mathematics, and alphabetical grades in 5-7 additional subjects including at least one Science subject

The characteristics of inconsistent reporting

In the next stage of the analyses, we estimate multinomial logistic regression models of inconsistent reporting in the two datasets. The outcome variables are whether young people under-report their GCSE results, over-report their GCSE results, or whether the results are consistent (i.e. their records match in both the self-reported and administrative datasets). The explanatory variables are derived from the main survey of the Millennium Cohort Study: parental social class (NS-SEC), sex, and administrative total points score (based on the interpolated score).

Table 8 presents the number of young people who over-report, under-report, or consistently report the number of GCSEs they achieved in the MCS, compared with the administrative data. Less than half of the young people consistently report the number of GCSEs they achieved, with fairly similar proportions of young people over- and under-reporting their qualifications. Table 9 presents a multinomial logistic regression model with a matching record as the base category. Compared with consistently reporting the number of GCSEs they achieved, young people with a higher overall points score were less likely to misreport the number of GCSEs attained (i.e. they were less likely to either under-report or over-report their GCSE qualifications than those with a lower overall points score). Females have significantly lower log odds of both under reporting and over reporting GCSE results. We do not observe an overall parental social class effect, although young people with parents in semi-routine and routine classes had significantly higher relative log odds of under-reporting their GCSEs, compared with young people with parents in higher professional classes.

Table 8: Over- and under-reporting of the number of GCSEs

	Frequency	Percentage	Adjusted Percentage
Under-reports	1,212	27.30	25.98
Matches	1,887	42.50	43.07
Over-report	1,341	30.20	30.96
Total	4,440	100.00	100.00

Table 9: Multinomial logistic regression of matching records of the number of GCSEs (base = Match)

	Under-Report			*Over-Report*		
	Log Odds		Standard Error	Log Odds		Standard Error
NPD interpolated scoring system	-0.04	***	(0.00)	-0.01	***	(0.00)
Sex						
Male	Ref			Ref		
Female	-0.41	***	(0.09)	-0.22	**	(0.08)
Parental NS-SEC						
1.1 Large Employers and Higher Managerial	0.17		(0.22)	0.10		(0.17)
1.2 Higher Professional	Ref.			Ref.		
2 Lower managerial and professional	0.07		(0.13)	0.02		(0.12)
3 Intermediate	0.26		(0.17)	-0.19		(0.16)
4 Small employers and own account	0.38	*	(0.16)	0.09		(0.16)
5 Lower Supervisory and Technical	0.29		(0.21)	-0.11		(0.20)
6 Semi-Routine	0.70	***	(0.18)	-0.15		(0.19)
7 Routine	0.44	*	(0.22)	-0.15		(0.17)
Constant	1.51	***	(0.19)	0.51	*	(0.22)
Observations	4,440					
McFadden's Adjusted R^2	0.05					
Cox-Snell R^2	0.11					
Nagelkerke R^2	0.13					
AIC	2.06					
BIC	-27924.34					

*Note: Self-reported data are from the Millennium Cohort Study (MCS) sweep 7 (SN8682). Administrative data are from the linked records from the National Pupil Database (SN8481). The models are adjusted for the complex sample design and non-response of the MCS. *p<.05 **p<.01 ***p<.001.*

Table 10 presents the number of young people who over-report, under-report, or consistently report the number of 'good' passes they achieve in their GCSEs. This is

not directly asked in the Millennium Cohort Study, but this is a derived measure which counts the grade information provided for each subject. This measure captures the extent to which young people accurately report their grades. Approximately half of the sample have matching records in the self-reported and administrative data. Only 9% under-report their grades, compared to a much greater proportion of young people over-reporting their grades. Table 11 presents the multinomial logistic regression model with matching records as the base category. As in the previous model, females had lower relative log odds than males of under-reporting and over-reporting the number of 'good' GCSE passes they achieved compared with having consistent records. Young people with higher overall points scores had lower relative log odds of over-reporting compared with having matching records, but the effect was not significant for under-reporting compared with matching. This is likely to be due to the much lower percentage of young people under-reporting their grades overall. There is no parental social class effect in either partition of the model.

Overall, there are differences in whether young people consistently or inconsistently self-report the number of GCSE qualifications and the number of 'good' passes they achieve, but there is no evidence of a systematic difference in a specific direction. We do not observe any characteristics that clearly differentiate those young people who under-report from those that over-report.

Table 10: Over- and under-reporting of the number of 'good' passes

	Frequency	Percentage	Adjusted Percentage
Under-reports	408	9.19	9.28
Matches	2,278	51.31	51.66
Over-report	1,754	39.50	39.06
Total	4,440	100.00	100.00

Table 11: Multinomial logistic regression of matching records of the number of 'good' passes (base = Match)

	Under-Report			*Over-Report*		
	Log Odds		Standard Error	Log Odds		Standard Error
NPD interpolated scoring system	0.00		(0.00)	-0.01	***	(0.00)
Sex						
Male	Ref.			Ref.		
Female	-0.29	*	(0.14)	-0.18	*	(0.07)
Parental NS-SEC						
1.1 Large Employers and Higher Managerial	0.28		(0.26)	0.08		(0.17)
1.2 Higher Professional	Ref.			Ref.		
2 Lower managerial and professional	-0.29		(0.18)	0.08		(0.13)
3 Intermediate	-0.17		(0.24)	-0.02		(0.14)
4 Small employers and own account	0.11		(0.20)	0.25		(0.14)
5 Lower Supervisory and Technical	-0.63		(0.32)	-0.10		(0.18)
6 Semi-Routine	0.16		(0.24)	-0.02		(0.16)
7 Routine	-0.51		(0.33)	0.08		(0.17)
Constant	-1.56	***	(0.24)	0.35		(0.18)
Observations	4,440					
McFadden's Adjusted R^2	0.00					
Cox-Snell R^2	0.02					
Nagelkerke R^2	0.02					
AIC	1.86					
BIC	-28824.32					

*Note: Self-reported data are from the Millennium Cohort Study (MCS) sweep 7 (SN8682). Administrative data are from the linked records from the National Pupil Database (SN8481). The models are adjusted for the complex sample design and non-response of the MCS. *p<.05 **p<.01 ***p<.001.*

The potential impact on empirical analyses

There are clear inconsistencies and discrepancies between the GCSE results self-reported in sweep 7 of the MCS, and the administrative records held in the NPD for the same individuals. This section explores the extent to which these inconsistencies have an impact on the substantive conclusions drawn from empirical analyses. We estimate a series of models which use the various GCSE summary measures as the outcome variables (Table 12 presents the descriptive statistics of these measures). Our explanatory variables are parental NS-SEC, parental education, and sex (see Table 3 for the descriptive statistics of the explanatory variables).

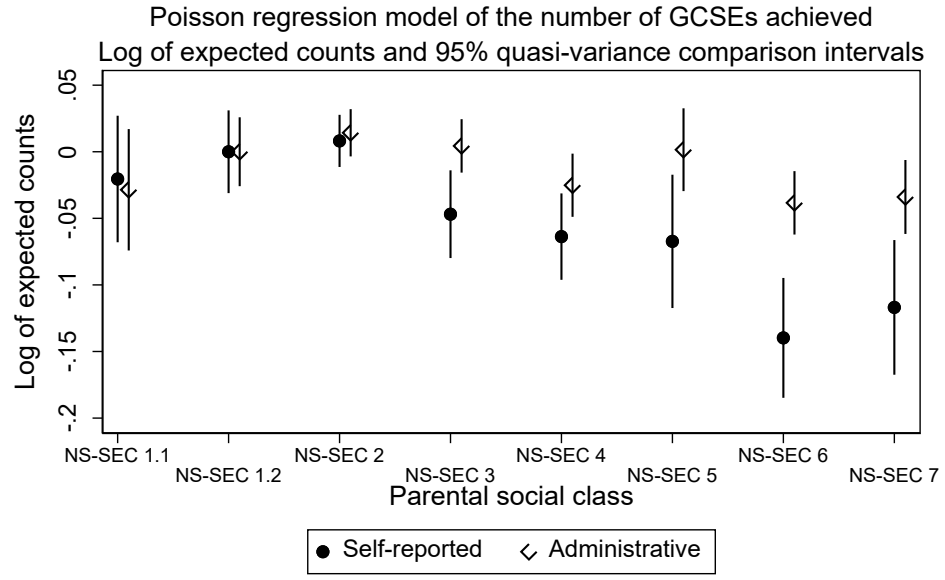
We present the regression modelling results in Figures 2-10, and the full model outputs can be found in the appendix. It is immediately clear that there is very little difference in the substantive results of the models analysing self-reported data and the model analysing administrative data. The most incongruent measure is the total number of GCSEs achieved, particularly for young people with parents in NS-SECs 5, 6, and 7. However, the magnitude of the difference is very small (<.1 in the log of expected count). For the blunter, more aggregated measures, the models are almost identical. For the more finely-grained measures, such as point scores, there are subtler differences in point estimates. Generally, we find that the self-reported data is systematically over-estimating social inequalities compared with the more conservative estimates from the administrative data, although the difference in the size of effects is very small. Across all measures the confidence intervals overlap between the models using self-reported or administrative data, and the overall substantive conclusions would remain the same if either data source was analysed. The similarity of the results in this particular analytical enterprise is fortunate. We advise that the similarity between results in the presence of two data sources that have discrepancies should not be assumed *a priori*. Whenever it is practicable the effects of inconsistencies and discrepancies between data sources should be investigated.

Table 12: Descriptive statistics of GCSE summary measures in the MCS and NPD

	Self-reported			Administrative		
	Frequency	**Percentage**	**Adjusted Percentage**	**Frequency**	**Percentage**	**Adjusted Percentage**
5+ 'good' passes						
Yes	3,578	80.59	80.40	3,360	75.68	75.72
No	862	19.41	19.60	1,080	24.32	24.28
5+ 'good' passes including English and Mathematics						
Yes	3,235	72.86	73.01	3,127	70.43	70.76
No	1,205	27.14	26.99	1,313	29.57	29.24
	Mean	**Standard deviation**		**Mean**	**Standard deviation**	
Number of GCSEs	8.83	2.56		8.92	1.88	
Number of 'good' passes	7.67	3.33		7.01	3.32	
Combined score	47.58	21.70		48.41	18.17	
Combined score (capped)	43.40	16.88		45.64	14.59	
Interpolated score	51.54	20.67		49.56	17.65	
Interpolated score (capped)	48.63	16.74		47.92	15.28	
Mean score	5.74	1.17		5.50	1.38	
Total	4,440			4,440		

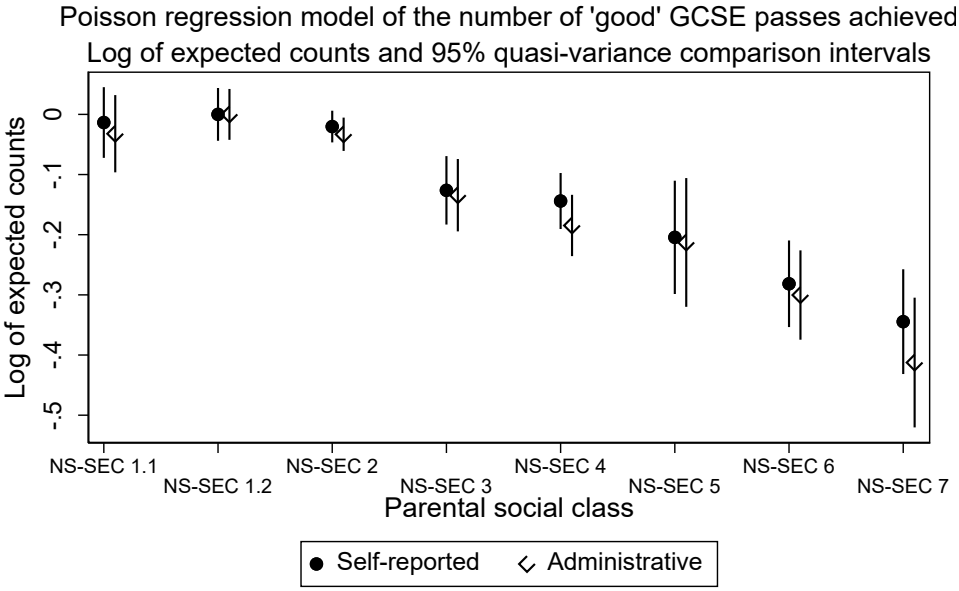
Note: we are unable to present the ranges of numeric measures due to statistical disclosure control.

The GCSE summary measures are described in Table 1.



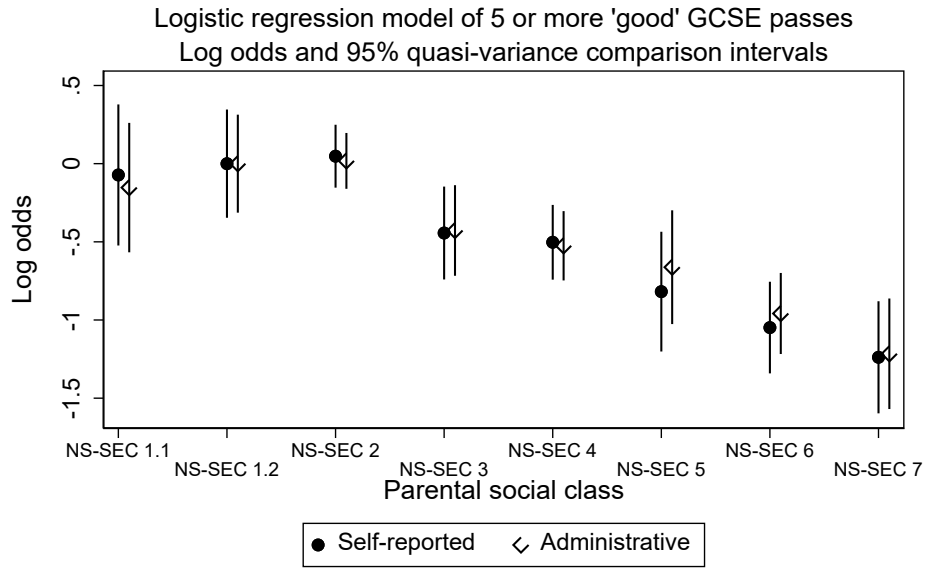
Self-reported data are from the Millennium Cohort Study (MCS) sweep 7 (SN8682).
Administrative data are from the linked records from the National Pupil Database (SN8481).
The models also include parental education level and sex, and adjust for the complex sample design of the MCS.
The models are estimated separately on the self-reported data and the administrative data, n=4,440.

Figure 2: Poisson regression model of the number of GCSE qualifications in the self-reported (MCS) and administrative (NPD) datasets



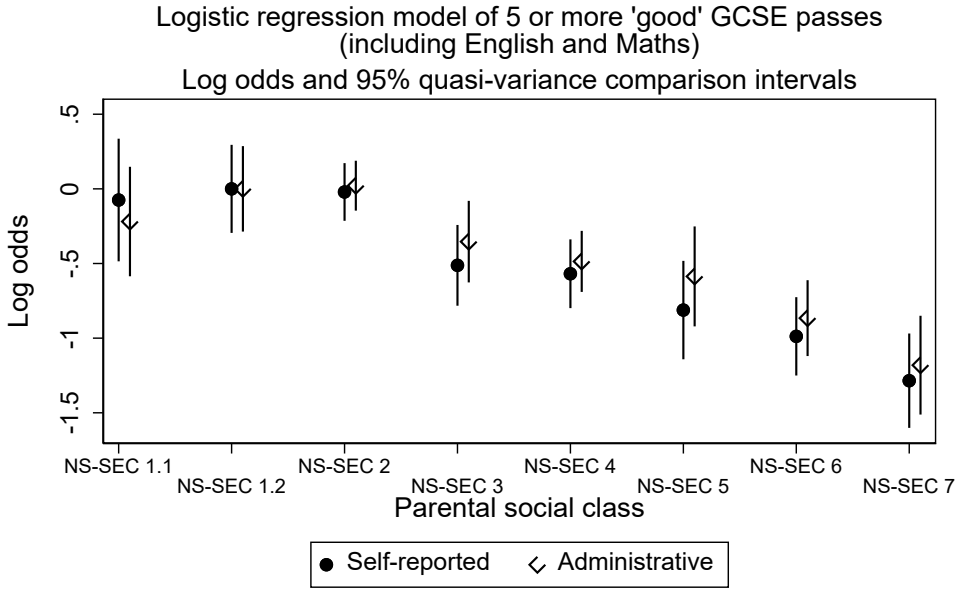
Self-reported data are from the Millennium Cohort Study (MCS) sweep 7 (SN8682).
Administrative data are from the linked records from the National Pupil Database (SN8481).
The models also include parental education level and sex, and adjust for the complex sample design of the MCS.
The models are estimated separately on the self-reported data and the administrative data, n=4,440.

Figure 3: Poisson regression model of the number of 'good' GCSE passes in the self-reported (MCS) and administrative (NPD) datasets



Self-reported data are from the Millennium Cohort Study (MCS) sweep 7 (SN8682).
Administrative data are from the linked records from the National Pupil Database (SN8481).
The models also include parental education level and sex, and adjust for the complex sample design of the MCS.
The models are estimated separately on the self-reported data and the administrative data, n=4,440.

Figure 4: Logistic regression model of 5 or more 'good' GCSE passes in the self-reported (MCS) and administrative (NPD) datasets



Self-reported data are from the Millennium Cohort Study (MCS) sweep 7 (SN8682).
Administrative data are from the linked records from the National Pupil Database (SN8481).
The models also include parental education level and sex, and adjust for the complex sample design of the MCS.
The models are estimated separately on the self-reported data and the administrative data, n=4,440.

Figure 5: Logistic regression models of 5 or more 'good' GCSE passes including English and Mathematics in the self-reported (MCS) and administrative (NPD) datasets

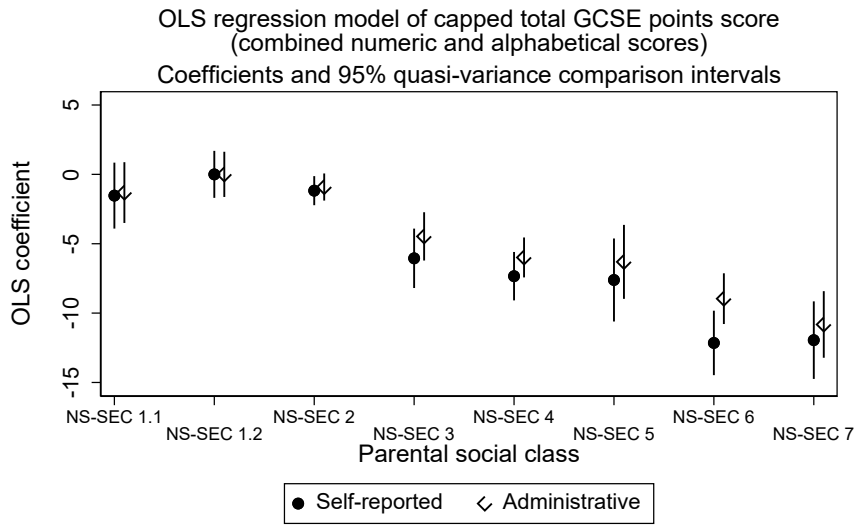


Figure 6: OLS regression model of total GCSE points score (combined numeric and alphabet scores) in the self-reported (MCS) and administrative (NPD) datasets

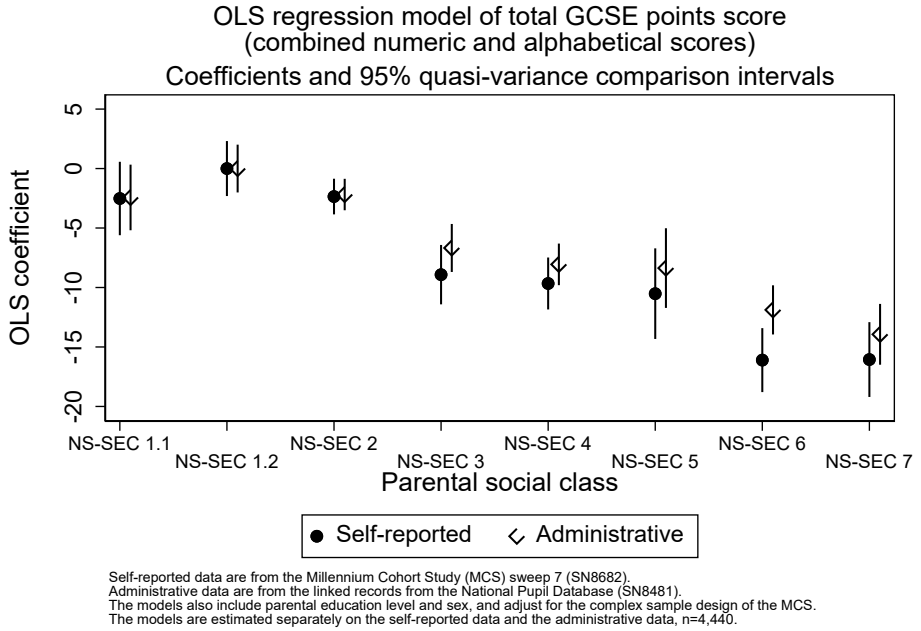
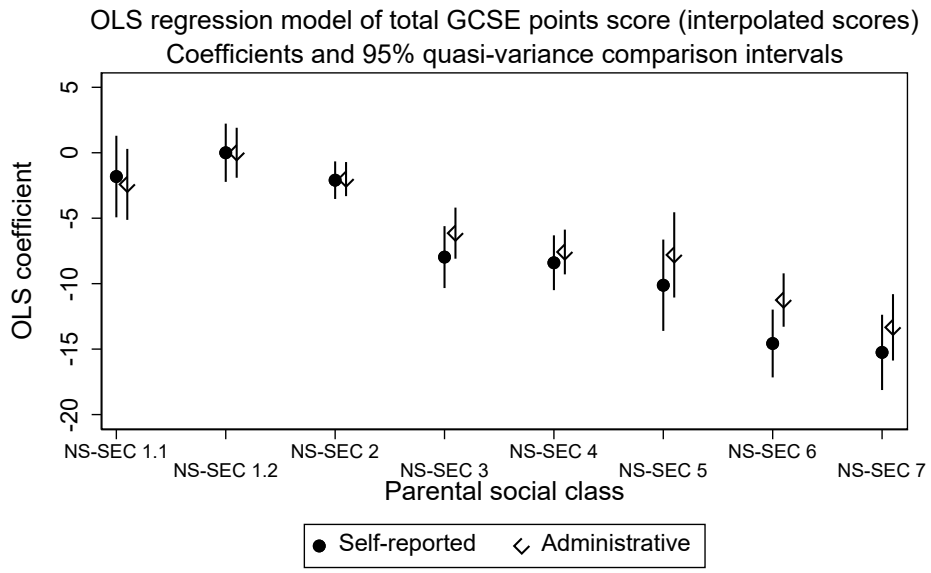
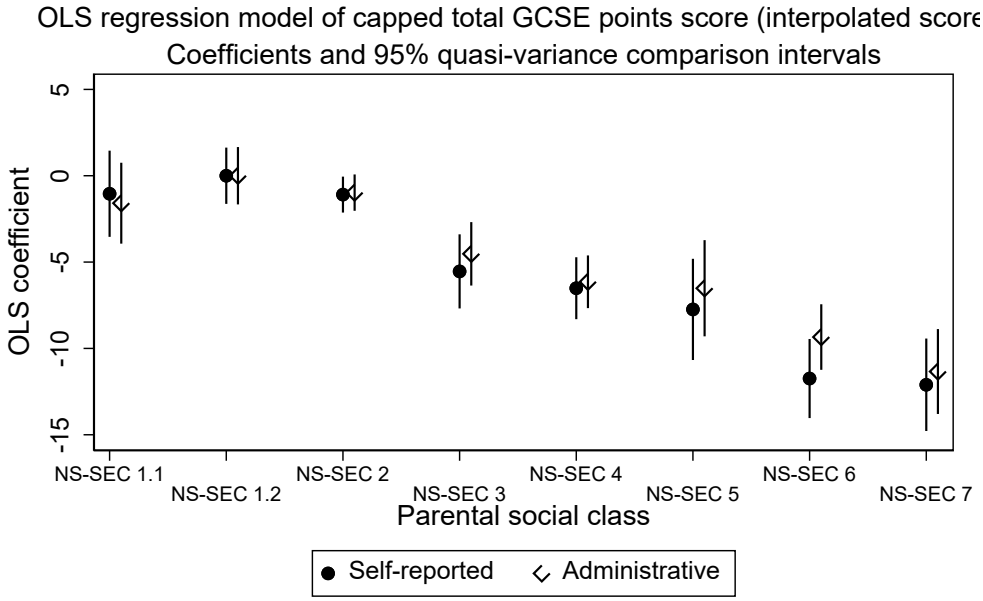


Figure 7: OLS regression model of capped total GCSE points score (combined numeric and alphabet scores) in the self-reported (MCS) and administrative (NPD) datasets



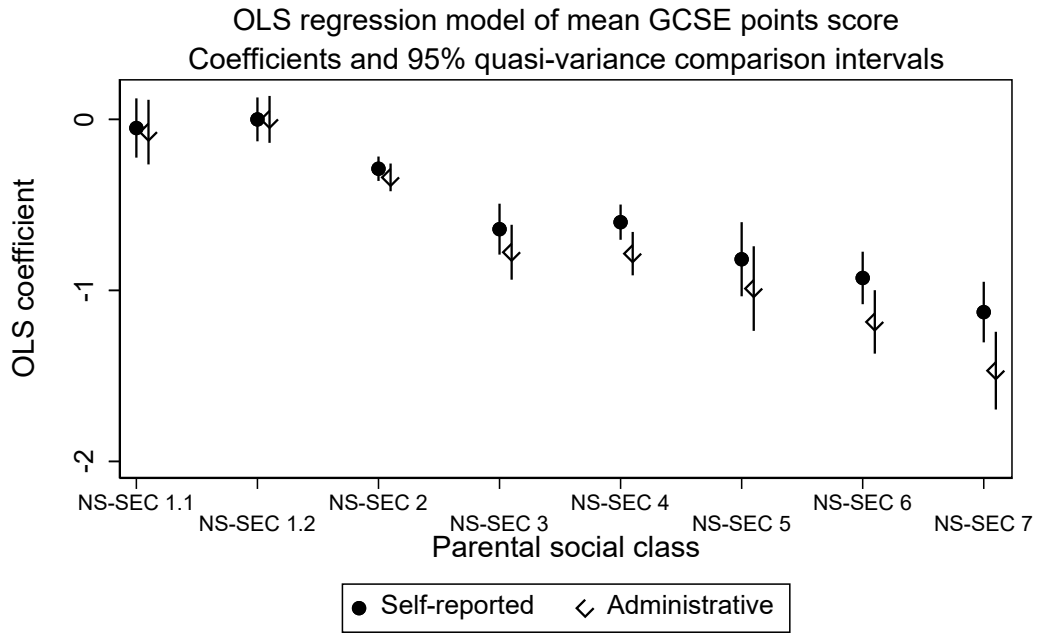
Self-reported data are from the Millennium Cohort Study (MCS) sweep 7 (SN8682).
Administrative data are from the linked records from the National Pupil Database (SN8481).
The models also include parental education level and sex, and adjust for the complex sample design of the MCS.
The models are estimated separately on the self-reported data and the administrative data, n=4,440.

Figure 8: OLS regression model of total GCSE points score (interpolated scores) in the self-reported (MCS) and administrative (NPD) datasets



Self-reported data are from the Millennium Cohort Study (MCS) sweep 7 (SN8682).
Administrative data are from the linked records from the National Pupil Database (SN8481).
The models also include parental education level and sex, and adjust for the complex sample design of the MCS.
The models are estimated separately on the self-reported data and the administrative data, n=4,440.

Figure 9: OLS regression model of capped total GCSE points score (interpolated scores) in the self-reported (MCS) and administrative (NPD) datasets



Self-reported data are from the Millennium Cohort Study (MCS) sweep 7 (SN8682).
Administrative data are from the linked records from the National Pupil Database (SN8481).
The models also include parental education level and sex, and adjust for the complex sample design of the MCS.
The models are estimated separately on the self-reported data and the administrative data, n=4,440.

Figure 10: OLS regression model of mean GCSE scores in the self-reported (MCS) and administrative (NPD) dataset

Conclusion

In the UK, there has been a promising advent of linked administrative records to large-scale, nationally representative social surveys which facilitate rich analyses at the individual-level. A major innovation that sets apart the Millennium Cohort Study (MCS) from the older British birth cohort studies is that data from administrative educational records on school qualifications have been linked to the study. The combination of NPD data and MCS survey data provides a powerful and unparalleled infrastructural resource for researchers studying school qualifications and educational inequalities in England, and for research in the broader academic fields such as social stratification and health.

In this project, we compared self-reported GCSE results in sweep 7 of the MCS with the GCSE results contained in administrative records in the NPD. Considering that the results were collected from the same individuals a year apart, we would expect to find highly congruent results. We compared a range of alternative summaries of GCSE results and found clear inconsistencies across all measures. There were greater matches for measures with lower resolution, such as the number of 'good' passes and the achievement of 5 or more 'good' passes. There were fewer direct matches for measures with higher resolution, such as those based on points scores. The number of matches is naturally related to degree of aggregation of these measures. This is a subtle but important demonstration of the potential impact of the inconsistencies when different approaches to summarising GCSE attainment are used. Nevertheless, correlations were generally high between self-reported and administrative measures. We found that females and people with higher overall scores were more likely to accurately report their results, but there was no pattern of under-reporting or over-reporting by males and people with lower overall scores. Despite the clear inconsistencies between the self-reported data and the administrative records, the impact on empirical analyses was fairly minimal. We found that the same substantive conclusions would be drawn regardless of the source of data for the outcome variable. We must stress, however, that this should not be assumed *a priori*, and may not be the case when answering other research questions.

On reflection, we consider that there is a general lack of precision when self-reporting results, which we conjecture is partially due to the way in which the data were collected. Despite great effort and consideration of alternative modes of data collection, the approach chosen has led to some issues and unintended consequences. In future collections of detailed school-level qualifications data, we would advise that using an electronic version of the qualifications grid, similar to that used in the data collection of the Youth Cohort Study of England and Wales, would be highly beneficial (see the questionnaire detailed on page 21 in Gray and Pattie, 1987). On consultation with the MCS team, we understand this was one of the potential approaches which was considered. We suggest that this data collection instrument is worthy of reconsideration by any forthcoming survey enterprises collecting GCSE data.

At GCSE level, there are 'core' subjects (e.g. English, Mathematics, and Science), 'foundation' subjects (e.g. computing, physical education, and citizenship), and schools must also offer at least one subject from each of the arts, design and technology, humanities, and modern foreign languages. Pupils may not sit GCSE examinations in all of these subjects, but they could form the basis of an initial list of GCSE qualifications which most pupils are likely to have been offered the opportunity to study. Using a pre-populated list of subjects which are likely to have been taken by many pupils would help young people recall each subject. It would further benefit researchers analysing these data to have core and foundation subjects readily available as variables within the dataset in a standardised format. This more structured approach to the collection of detailed educational measures is likely to improve the quality of the data collected via a self-reporting instrument.

Considering the large discrepancies in mentioning subjects more than once, and the much larger proportion of double awards flagged in the self-reported data compared with the administrative records, we would recommend that data collectors validate the self-reported data. We are aware that data collectors intend to ask the cohort members to verify their GCSE results in sweep 8 (age 23) of the MCS. In future cohort studies, we would recommend that data collectors consider asking interviewers to undertake a contemporaneous check of an official examination transcript of the young person's GCSE results and provide a flag in the data for analysts to ascertain if the results have

been independently verified or not. Although we appreciate that survey data collectors are required to navigate a delicate balance between the ideal data collection process and the burden placed on research participants. We understand that a contemporaneous check was carefully considered for the MCS survey data collection but it was ruled out after considering the practical drawbacks. In future collections of detailed school-level qualifications data, collectors may also consider using innovative data collection approaches, such as requesting a photograph of the exam certificate around results time. A similar approach has been trialled in *Understanding Society - The United Kingdom Household Longitudinal Study,* which collects information from newborn babies' red books (Benzeval et al., 2024).

The MCS is an important data resource for educational research, and it is uncommon for a social survey to collect detailed, subject-specific qualifications data from its respondents. The ease of access to the self-reported data may compel data analysts to use these data. We have demonstrated that there are a sufficient number of obvious inconsistencies in the data. There is also a lack of accompanying documentation for the self-reported educational data. Therefore, we would caution researchers from using these data uncritically. To accompany this paper, we are providing access to the research code required to process data from both the MCS and the NPD (see https://osf.io/gs6m2/).

We envisaged that the administrative data would provide a reliable and consistent source of information. However, we have demonstrated that there are also inconsistencies in the linked NPD records, e.g. some GCSE subjects, most notably the core subject of English Language, are not deposited in the pupil-level dataset. We have discovered that neither the survey records nor the administrative records provide a 'research ready' data source for analyses. We issue the strong recommendation that data analysts use the raw NPD data (i.e. the raw subject and qualification codes, in contrast to the readily deposited 'pupil level' records[2]). The raw NPD data also requires suitable data wrangling to prepare the records and we recommend that researchers draw on the research code that we have deposited.

[2] This is the pupil-level file in the linked dataset: MCS_CM_NPD_KS4_PUPIL_2017 (see page 8 in Rihal and Gomes, 2021).

Appendix – Full regression model output

Table A1: Poisson regression model - Number of GCSEs achieved

	Self-reported data					Administrative data				
	Log of expected count	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error	Log of expected count	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error
Parental NS-SEC										
1.1 Large Employers and Higher Managerial	-0.02	(0.02)	0.02	-0.19	(0.22)	-0.03	(0.02)	0.02	-0.25	(0.18)
1.2 Higher Professional	0.00	(.)	0.01	0.00	(.)	0.00	(.)	0.01	0.00	(.)
2 Lower Managerial and Professional	0.01	(0.02)	0.01	0.08	(0.15)	0.01	(0.01)	0.01	0.13	(0.12)
3 Intermediate	-0.05*	(0.02)	0.01	-0.42*	(0.18)	0.00	(0.01)	0.01	0.04	(0.13)
4 Small Employers and Own Account	-0.06**	(0.02)	0.01	-0.57**	(0.18)	-0.03	(0.01)	0.01	-0.22	(0.13)
5 Lower Supervisory and Technical	-0.07**	(0.03)	0.02	-0.60**	(0.22)	0.00	(0.02)	0.01	0.01	(0.17)
6 Semi-Routine	-0.14***	(0.02)	0.02	-1.20***	(0.20)	-0.04*	(0.02)	0.01	-0.34*	(0.15)
7 Routine	-0.12***	(0.03)	0.02	-1.01***	(0.23)	-0.03	(0.02)	0.01	-0.30	(0.16)
Parental Education										
No degree	0.00	(.)		0.00	(.)	0.00	(.)		0.00	(.)
Degree	0.02	(0.01)		0.20	(0.11)	-0.00	(0.01)		-0.04	(0.09)
Sex										
Male	0.00	(.)		0.00	(.)	0.00	(.)		0.00	(.)
Female	0.05***	(0.01)		0.42***	(0.09)	0.02**	(0.01)		0.20**	(0.06)
Constant	2.19***	(0.01)				2.18***	(0.01)			
Observations	4,440					4,440				
McFadden's Adjusted R^2	0.00					-0.00				
Cox-Snell R^2	0.03					0.00				
Nagelkerke R^2	0.03					0.00				

	Self-reported data					Administrative data				
	Log of expected count	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error	Log of expected count	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error
AIC	4.83					4.49				
BIC	-15754.35					-17278.79				

Table A2: Poisson regression model - Number of 'good' GCSE passes

	Self-reported data					Administrative data				
	Log of expected count	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error	Log of expected count	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error
Parental NS-SEC										
1.1 Large Employers and Higher Managerial	-0.01	(0.03)	0.03	-0.11	(0.27)	-0.03	(0.03)	0.03	-0.25	(0.25)
1.2 Higher Professional	0.00	(.)	0.02	0.00	(.)	0.00	(.)	0.02	0.00	(.)
2 Lower Managerial and Professional	-0.02	(0.02)	0.01	-0.17	(0.19)	-0.03	(0.02)	0.01	-0.26	(0.18)
3 Intermediate	-0.13***	(0.03)	0.03	-1.00***	(0.23)	-0.13***	(0.03)	0.03	-0.99***	(0.22)
4 Small Employers and Own Account	-0.14***	(0.03)	0.02	-1.13***	(0.22)	-0.18***	(0.03)	0.02	-1.32***	(0.22)
5 Lower Supervisory and Technical	-0.20***	(0.05)	0.04	-1.56***	(0.34)	-0.21***	(0.05)	0.05	-1.51***	(0.34)
6 Semi-Routine	-0.28***	(0.04)	0.03	-2.07***	(0.26)	-0.30***	(0.04)	0.03	-2.04***	(0.24)
7 Routine	-0.34***	(0.04)	0.04	-2.46***	(0.27)	-0.41***	(0.05)	0.05	-2.66***	(0.29)
Parental Education										
No degree	-0.07***	(0.02)		-0.54***	(0.14)	-0.06**	(0.02)		-0.43**	(0.14)
Degree	0.00	(.)		0.00	(.)	0.00	(.)		0.00	(.)
Sex										
Male	-0.11***	(0.01)		-0.82***	(0.11)	-0.13***	(0.02)		-0.94***	(0.11)
Female	0.00	(.)		0.00	(.)	0.00	(.)		0.00	(.)
Constant	2.24***	(0.02)				2.17***	(0.02)			

	Self-reported data					Administrative data				
	Log of expected count	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error	Log of expected count	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error
Observations	4,440					4,440				
McFadden's Adjusted R^2	0.02					0.02				
Cox-Snell R^2	0.10					0.12				
Nagelkerke R^2	0.10					0.12				
AIC	5.46					5.53				
BIC	-12955.58					-12664.16				

Table A3: Logistic regression model - 5 or more 'good' GCSE passes

	Self-reported data					Administrative data				
	Log Odds	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error	Log Odds	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error
Parental NS-SEC										
1.1 Large Employers and Higher Managerial	-0.07	(0.25)	0.20	-0.01	(0.03)	-0.15	(0.22)	0.18	-0.02	(0.03)
1.2 Higher Professional	0.00	(.)	0.15	0.00	(.)	0.00	(.)	0.14	0.00	(.)
2 Lower Managerial and Professional	0.05	(0.18)	0.09	0.01	(0.02)	0.02	(0.16)	0.08	0.00	(0.02)
3 Intermediate	-0.44*	(0.20)	0.13	-0.06*	(0.03)	-0.43*	(0.18)	0.13	-0.07*	(0.03)
4 Small Employers and Own Account	-0.50**	(0.19)	0.11	-0.07**	(0.03)	-0.53**	(0.17)	0.10	-0.09**	(0.03)
5 Lower Supervisory and Technical	-0.82***	(0.23)	0.17	-0.13***	(0.04)	-0.66**	(0.21)	0.16	-0.12**	(0.04)
6 Semi-Routine	-1.05***	(0.21)	0.13	-0.18***	(0.03)	-0.96***	(0.18)	0.11	-0.18***	(0.03)
7 Routine	-1.24***	(0.21)	0.16	-0.22***	(0.04)	-1.22***	(0.21)	0.16	-0.24***	(0.04)
Parental Education										
No degree	-0.25	(0.13)		-0.04*	(0.02)	-0.25*	(0.12)		-0.04*	(0.02)
Degree	0.00	(.)		0.00	(.)	0.00	(.)		0.00	(.)
Sex										
Male	-0.55***	(0.10)		-0.08***	(0.01)	-0.56***	(0.08)		-0.10***	(0.01)
Female	0.00	(.)		0.00	(.)	0.00	(.)		0.00	(.)
Constant	2.29***	(0.19)				2.00***	(0.17)			
Observations	4,440					4,440				
McFadden's Adjusted R^2	0.03					0.03				

	Self-reported data					Administrative data				
	Log Odds	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error	Log Odds	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error
Cox-Snell R^2	0.04					0.04				
Nagelkerke R^2	0.06					0.06				
AIC	0.95					1.07				
BIC	-32976.88					-32444.59				

Table A4: Logistic regression model - 5 or more 'good' GCSE passes (including English and Mathematics)

	Self-reported data					Administrative data				
	Log Odds	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error	Log Odds	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error
Parental NS-SEC										
1.1 Large Employers and Higher Managerial	-0.07	(0.22)	0.18	-0.01	(0.04)	-0.22	(0.20)	0.16	-0.04	(0.04)
1.2 Higher Professional	0.00	(.)	0.13	0.00	(.)	0.00	(.)	0.13	0.00	(.)
2 Lower Managerial and Professional	-0.02	(0.15)	0.09	-0.00	(0.02)	0.02	(0.14)	0.07	0.00	(0.02)
3 Intermediate	-0.51**	(0.17)	0.12	-0.09**	(0.03)	-0.35*	(0.17)	0.12	-0.07*	(0.03)
4 Small Employers and Own Account	-0.57***	(0.17)	0.10	-0.10***	(0.03)	-0.49**	(0.16)	0.09	-0.10**	(0.03)
5 Lower Supervisory and Technical	-0.81***	(0.20)	0.15	-0.16***	(0.04)	-0.59**	(0.19)	0.15	-0.12**	(0.04)
6 Semi-Routine	-0.99***	(0.18)	0.12	-0.20***	(0.04)	-0.87***	(0.17)	0.11	-0.18***	(0.04)
7 Routine	-1.28***	(0.19)	0.14	-0.27***	(0.04)	-1.18***	(0.20)	0.15	-0.26***	(0.04)
Parental Education										
No degree	-0.14	(0.12)		-0.03	(0.02)	-0.10	(0.11)		-0.02	(0.02)
Degree	0.00	(.)		0.00	(.)	0.00	(.)		0.00	(.)
Sex										
Male	-0.38***	(0.08)		-0.07***	(0.02)	-0.42***	(0.08)		-0.08***	(0.02)
Female	0.00	(.)		0.00	(.)	0.00	(.)		0.00	(.)

	Self-reported data					Administrative data				
	Log Odds	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error	Log Odds	Standard Error	QV Standard Error	Average Marginal Effects	Standard Error
Constant	1.70***	(0.16)				1.50***	(0.16)			
Observations	4,440					4,440				
McFadden's Adjusted R^2	0.03					0.02				
Cox-Snell R^2	0.04					0.03				
Nagelkerke R^2	0.05					0.05				
AIC	1.14					1.19				
BIC	-32153.90					-31936.48				

Table A5: OLS regression model - Total points score (combined numeric and alphabet scores)

	Self-reported data			Administrative data		
	OLS Coefficient	Standard Error	QV Standard Error	OLS Coefficient	Standard Error	QV Standard Error
Parental NS-SEC						
1.1 Large Employers and Higher Managerial	-2.52	(1.73)	1.36	-2.43	(1.43)	1.22
1.2 Higher Professional	0.00	(.)	1.02	0.00	(.)	0.89
2 Lower Managerial and Professional	-2.35	(1.24)	0.66	-2.18*	(1.08)	0.58
3 Intermediate	-8.92***	(1.43)	1.10	-6.68***	(1.21)	0.89
4 Small Employers and Own Account	-9.66***	(1.45)	0.97	-8.06***	(1.21)	0.77
5 Lower Supervisory and Technical	-10.52***	(2.01)	1.68	-8.37***	(1.76)	1.48
6 Semi-Routine	-16.10***	(1.50)	1.19	-11.88***	(1.27)	0.91
7 Routine	-16.06***	(1.72)	1.39	-13.94***	(1.46)	1.13
Parental Education						
No degree	-4.70***	(0.90)		-3.60***	(0.80)	
Degree	0.00	(.)		0.00	(.)	
Sex						
Male	-5.48***	(0.74)		-4.94***	(0.61)	
Female	0.00	(.)		0.00	(.)	
Constant	60.47***	(1.28)		58.79***	(1.02)	
Observations	4,440			4,440		
Adjusted R^2	0.09			0.08		
AIC	8.90			8.56		
BIC	2317.73			786.32		

Table A6: OLS regression model - Capped total points score (combined numeric and alphabet scores)

	Self-reported data			Administrative data		
	OLS Coefficient	Standard Error	QV Standard Error	OLS Coefficient	Standard Error	QV Standard Error
Parental NS-SEC						
1.1 Large Employers and Higher Managerial	-1.53	(1.30)	1.05	-1.31	(1.16)	0.97
1.2 Higher Professional	0.00	(.)	0.75	0.00	(.)	0.72
2 Lower Managerial and Professional	-1.17	(0.87)	0.46	-0.91	(0.85)	0.43
3 Intermediate	-6.05***	(1.16)	0.95	-4.47***	(1.02)	0.77
4 Small Employers and Own Account	-7.34***	(1.13)	0.77	-5.99***	(0.98)	0.64
5 Lower Supervisory and Technical	-7.61***	(1.54)	1.32	-6.31***	(1.42)	1.18
6 Semi-Routine	-12.15***	(1.23)	1.03	-8.96***	(1.06)	0.81
7 Routine	-11.95***	(1.47)	1.24	-10.82***	(1.31)	1.06
Parental Education						
No degree	-2.54***	(0.66)		-2.01**	(0.60)	
Degree	0.00	(.)		0.00	(.)	
Sex						
Male	-4.33***	(0.59)		-3.86***	(0.47)	
Female	0.00	(.)		0.00	(.)	
Constant	52.20***	(0.93)		52.71***	(0.83)	
Observations	4,440			4,440		
Adjusted R^2	0.08			0.07		
AIC	8.41			8.13		
BIC	118.72			-1130.13		

Table A7: OLS regression model - Total points score (interpolated scores)

	Self-reported data			Administrative data		
	OLS Coefficient	Standard Error	QV Standard Error	OLS Coefficient	Standard Error	QV Standard Error
Parental NS-SEC						
1.1 Large Employers and Higher Managerial	-1.82	(1.72)	1.38	-2.42	(1.39)	1.20
1.2 Higher Professional	0.00	(.)	0.98	0.00	(.)	0.84
2 Lower Managerial and Professional	-2.10	(1.20)	0.63	-2.01	(1.04)	0.58
3 Intermediate	-7.97***	(1.39)	1.05	-6.14***	(1.16)	0.86
4 Small Employers and Own Account	-8.40***	(1.37)	0.93	-7.58***	(1.17)	0.76
5 Lower Supervisory and Technical	-10.12***	(1.84)	1.54	-7.80***	(1.71)	1.44
6 Semi-Routine	-14.57***	(1.46)	1.15	-11.25***	(1.23)	0.90
7 Routine	-15.25***	(1.58)	1.27	-13.34***	(1.43)	1.12
Parental Education						
No degree	-4.57***	(0.87)		-3.32***	(0.77)	
Degree	0.00	(.)		0.00	(.)	
Sex						
Male	-5.05***	(0.69)		-4.77***	(0.58)	
Female	0.00	(.)		0.00	(.)	
Constant	63.42***	(1.28)		59.31***	(0.98)	
Observations	4,440			4,440		
Adjusted R^2	0.08			0.08		
AIC	8.81			8.50		
BIC	1912.74			545.93		

Table A8: OLS regression model - Capped total points score (interpolated scores)

	Self-reported data			Administrative data		
	OLS Coefficient	Standard Error	QV Standard Error	OLS Coefficient	Standard Error	QV Standard Error
Parental NS-SEC						
1.1 Large Employers and Higher Managerial	-1.04	(1.35)	1.10	-1.59	(1.22)	1.04
1.2 Higher Professional	0.00	(.)	0.72	0.00	(.)	0.73
2 Lower Managerial and Professional	-1.09	(0.86)	0.46	-0.98	(0.88)	0.47
3 Intermediate	-5.54***	(1.16)	0.95	-4.52***	(1.05)	0.81
4 Small Employers and Own Account	-6.51***	(1.08)	0.79	-6.14***	(1.02)	0.67
5 Lower Supervisory and Technical	-7.74***	(1.49)	1.30	-6.52***	(1.48)	1.23
6 Semi-Routine	-11.74***	(1.21)	1.01	-9.34***	(1.10)	0.84
7 Routine	-12.10***	(1.40)	1.18	-11.34***	(1.35)	1.09
Parental Education						
No degree	-2.77***	(0.67)		-2.09**	(0.65)	
Degree	0.00	(.)		0.00	(.)	
Sex						
Male	-4.21***	(0.57)		-3.99***	(0.49)	
Female	0.00	(.)		0.00	(.)	
Constant	57.26***	(0.93)		55.26***	(0.86)	
Observations	4,440			4,440		
Adjusted R^2	0.08			0.07		
AIC	8.39			8.22		
BIC	51.26			-707.66		

Table A9: OLS regression model - Mean points score (based on interpolated scores)

	Self-reported data			Administrative data		
	OLS Coefficient	Standard Error	QV Standard Error	OLS Coefficient	Standard Error	QV Standard Error
Parental NS-SEC						
1.1 Large Employers and Higher Managerial	-0.05	(0.09)	0.08	-0.07	(0.10)	0.08
1.2 Higher Professional	0.00	(.)	0.06	0.00	(.)	0.06
2 Lower Managerial and Professional	-0.29***	(0.07)	0.03	-0.34***	(0.07)	0.04
3 Intermediate	-0.64***	(0.08)	0.07	-0.78***	(0.09)	0.07
4 Small Employers and Own Account	-0.60***	(0.07)	0.05	-0.79***	(0.09)	0.06
5 Lower Supervisory and Technical	-0.82***	(0.11)	0.10	-0.99***	(0.13)	0.11
6 Semi-Routine	-0.93***	(0.09)	0.07	-1.18***	(0.10)	0.08
7 Routine	-1.13***	(0.10)	0.08	-1.47***	(0.12)	0.10
Parental Education						
No degree	-0.39***	(0.05)		-0.42***	(0.06)	
Degree	0.00	(.)		0.00	(.)	
Sex						
Male	-0.30***	(0.04)		-0.42***	(0.05)	
Female	0.00	(.)		0.00	(.)	
Constant	6.63***	(0.06)		6.62***	(0.07)	
Observations	4,440			4,440		
Adjusted R^2	0.12			0.15		
AIC	3.02			3.33		
BIC	-23780.05			-22441.62		

Data citation

University of London, Institute of Education, Centre for Longitudinal Studies. (2023). Millennium Cohort Study: Age 14, Sweep 6, 2015. [data collection]. 7th Edition. UK Data Service. SN: 8156, DOI: http://doi.org/10.5255/UKDA-SN-8156-7

University of London, Institute of Education, Centre for Longitudinal Studies. (2023). Millennium Cohort Study: Age 17, Sweep 7, 2018. [data collection]. 2nd Edition. UK Data Service. SN: 8682, DOI: http://doi.org/10.5255/UKDA-SN-8682-2

University of London, Institute of Education, Centre for Longitudinal Studies. (2023). Millennium Cohort Study: Sweeps 1-7, 2001-2018: Longitudinal Family File. [data collection]. 4th Edition. UK Data Service. SN: 8172, DOI: http://doi.org/10.5255/UKDA-SN-8172-4

University College London, UCL Institute of Education, Centre for Longitudinal Studies, Department for Education. (2021). Millennium Cohort Study: Linked Education Administrative Datasets (National Pupil Database), England: Secure Access. [data collection]. 2nd Edition. UK Data Service. SN: 8481, http://doi.org/10.5255/UKDA-SN-8481-2

References

Administrative Data Taskforce 2012. The UK Administrative Data Research Network: Improving access for research and policy. London: ESRC.

Adriaans, J., P. Valet and S. Liebig 2020. Comparing administrative and survey data: Is information on education from administrative records of the German Institute for Employment Research consistent with survey self-reports? *Quality & Quantity,* 54: 3-25.

Anders, J., F. Green, M. Henderson and G. Henseke 2024. Private school pupils' performance in GCSEs (and IGCSEs). *Cambridge Journal of Education,* 54: 795-813.

Benzeval, M., E. Aguirre, T. Al Baghal and L. Mitchell 2024. Obtaining measurement data from the 'red book'. *In:* Vine, J., E. Aguirre, T. Al Baghal, M. Benzeval, J. Burton, C. Butler, H. Chung, M. Couper, A. Coutrot, L. Delaney, C. Fowler, A. Jäckle, M. Kumari, M. Lieutaud, K. L. Mansfield, L. Mitchell, V. Parutis, J. Payne, G. Popli, A. K. Przybylski, S. Raj, A. Ratcliffe, A. R. Soetevent, H. Spiers, G. J. Van Den Berg, L. VoorintholtandS. Wang (eds.) *Understanding Society Innovation Panel wave 16: Results from methodological experiments and new data. Understanding Society Working Paper Series No. 2024 – 11 November 2024.* Essex: Understanding Society.

Card, D., R. Chetty, M. S. Feldstein and E. Saez 2010. Expanding access to administrative data for research in the United States. *American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas, Available at SSRN: https://ssrn.com/abstract=1888586 or http://dx.doi.org/10.2139/ssrn.1888586*.

Carpenter, J. and M. Kenward 2013. *Multiple imputation and its applications,* Chichester, John Wiley & Sons.

Connelly, R., V. Gayle and P. S. Lambert 2016a. A review of educational attainment measures for social survey research. *Methodological innovations,* 9: 205979911663800.

Connelly, R. and L. Platt 2014. Cohort profile: UK Millennium Cohort Study (MCS). *Int J Epidemiol,* 43: 1719-25.

Connelly, R., C. J. Playford, V. Gayle and C. Dibben 2016b. The role of administrative data in the big data revolution in social science research. *Social Science Research*.

Connelly, R., C. J. Playford, V. Gayle and C. Dibben 2016c. The role of administrative data in the big data revolution in social science research. *Soc Sci Res,* 59: 1-12.

Elliot Major, L. and S. Parsons 2022. The forgotten fifth: Examining the early education trajectories of teenagers who fall below the expected standards in GCSE English language and maths examinations at age 16. CLS working paper nummber 2022/6. London: UCL Centre for Longitudinal Studies.

Gayle, V. and R. Connelly 2022. The Stark realities of reproducible statistically orientated sociological research: Some newer rules of the sociological method. *Methodological Innovations,* 15: 207-221.

Gill, T. 2016. Uptake of Level 2 qualifications in English schools 2015. In Cambridge Assessment Statistics Report Series No.103. . Cambridge: Cambridge Assessment.

Goerge, R. M. and B. J. Lee 2001. Matching and cleaning administrative data. *In:* Citro, C. F., R. A. MoffittandM. Van Ploeg (eds.) *Studies of Higher Population: Data Collection and Research Issues.* Washington D.C.: National Academies Press.

Gould, W. W. 2006. Stata tip 35: Detecting whether data have changed. *Stata Journal,* 6: 428–429.

Gray, J. and C. Pattie 1987. An introduction to the Youth Cohort Study: Codebook for Cohort 1 Sweep 1. *In:* Courtenay, G.andS. Elder (eds.). University of Sheffield, Division of Education: Sheffield.

Hansen, K. (ed.) 2014. *Millennium Cohort Study: A guide to the datasets (eighth edition). First, second, third, fourth and fifth surveys,* London: Centre for Longitudinal Studies.

Harron, K., C. Dibben, J. Boyd, A. Hjern, M. Azimaee, M. L. Barreto and H. Goldstein 2017. Challenges in administrative data linkage for research. *Big Data Soc,* 4: 1-12.

Jay, M. A., L. Mc Grath-Lone and R. Gilbert 2019. Data Resource: the National Pupil Database (NPD). *International journal of population data science,* 4: 1101-1101.

Jerrim, J., P. Parker and S. Nikki 2019. Bullshitters. Who are they and what do we know about their lives? IZA DP No. 12282. Bonn, Germany: IZA Institute of Labor Economics.

Jones, K. H., S. Heys, K. S. Tingay, P. Jackson and C. Dibben 2018. The Good, the Bad, the Clunky: Improving the Use of Administrative Data for Research. *International journal of population data science,* 4: 587.

Kuncel, N. R., M. Credé and L. L. Thomas 2005. The Validity of Self-Reported Grade Point Averages, Class Ranks, and Test Scores: A Meta-Analysis and Review of the Literature. *Review of Educational Research,* 75: 63-82.

Leckie, G. and H. Goldstein 2009. The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society. Series A (Statistics in Society),* 172: 835-851.

Leckie, G. and H. Goldstein 2017. The evolution of school league tables in England 1992–2016: 'Contextual value-added', 'expected progress' and 'progress 8'. *British Educational Research Journal,* 43: 193-212.

Ofqual. 2018. *Grading new GCSEs* [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/719124/Grading_new_GCSEs25.6.2018.pdf [Accessed 13/06/2025].

Playford, C. J., V. Gayle, R. Connelly and A. J. G. Gray 2016. Administrative social science data: The challenge of reproducible research. *Big data & society,* 3: 1-13.

Rihal, S. and D. Gomes 2021. Millennium Cohort Study: A guide to the linked education administrative datasets (2nd edition). London: UCL Centre for Longitudinal Studies.

Rose, D. and D. J. Pevalin 2003. *A researcher's guide to the national statistics socio-economic classification,* London, SAGE.

Smith, K. and H. Joshi 2002. The Millennium Cohort Study. *Population Trends,* 107: 30-34.

Sticca, F. a.-O., T. Goetz, M. Bieg, N. C. Hall, F. Eberle and L. Haag 2017. Examining the accuracy of students' self-reported academic grades from a correlational and a discrepancy perspective: Evidence from a longitudinal study.

University College London, Ucl Institute of Education, Centre for Longitudinal Studies and D. F. Education 2021. Millennium Cohort Study: Linked Education Administrative Datasets (National Pupil Database), England: Secure Access. [data collection]. 2nd Edition. UK Data Service. SN: 8481, DOI: http://doi.org/10.5255/UKDA-SN-8481-2.

University of London, Institute of Education and C. F. L. Studies 2023. Millennium Cohort Study: Age 17, Sweep 7, 2018. [data collection]. 2nd Edition. UK Data Service. SN: 8682, DOI: http://doi.org/10.5255/UKDA-SN-8682-2.

Wadsworth, M. E. and J. Bynner (eds.) 2011. *A companion to life course studies: The social and historical context of the British birth cohort studies,* London: Routledge.

Walker, I. and T. Gamble 2023. Active travel to school: a longitudinal millennium cohort study of schooling outcomes. *BMJ Open,* 13: e068388.

Woollard, M. 2014. Administrative data: Problems and benefits. A perspective from the United Kingdom. *In:* A., D., N. D.andS. G. (eds.) *Facing the future: European research infrastructures for the Humanities and Social Sciences.* Berlin: SCIVERO Verlag.

Yang, M. and G. Woodhouse 2001. Progress from GCSE to A and AS Level: Institutional and Gender Differences, and Trends over Time. *British Educational Research Journal,* 27: 245-267.