

Measuring parental income using administrative data. What is the best proxy available?

John Jerrim

To cite this article: John Jerrim (03 Nov 2023): Measuring parental income using administrative data. What is the best proxy available?, Research Papers in Education, DOI: [10.1080/02671522.2023.2271930](https://doi.org/10.1080/02671522.2023.2271930)

To link to this article: <https://doi.org/10.1080/02671522.2023.2271930>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 03 Nov 2023.



[Submit your article to this journal](#)



Article views: 318



[View related articles](#)



[View Crossmark data](#)

Measuring parental income using administrative data. What is the best proxy available?

John Jerrim

Department of Social Science, UCL Institute of Education, London, United Kingdom of Great Britain and Northern Ireland

ABSTRACT

Administrative data are increasingly being used to study inequalities in education. Yet a well-known difficulty with such resources is the limited information they hold. A commonly used proxy for children coming from a low-income background is their eligibility for free school meals, yet this is likely to be of little use in measuring academic achievement amongst middle and high-income groups. This study adds to the literature by showing how eligibility for free school meals – averaged over the time a child has spent at school – is the best available proxy for low income during childhood. In contrast, creating a continuous index combining free school meal eligibility with information on the neighbourhood in which they live represents the best way of comparing educational outcomes across children from low, average and high-income backgrounds.

ARTICLE HISTORY

Received 18 April 2022
Accepted 8 September 2023

KEYWORDS


Administrative data; proxy measures; income-achievement gaps; permanent income

1. Introduction

There has long been interest in the social and medical sciences in socio-economic inequalities in education and health (Broer, Bai, and Fonseca 2019; Lago et al. 2018). Numerous studies have documented how large socio-economic disparities across several dimensions emerge early in life (UNICEF Office of Research 2018) and continue to influence educational achievement (Crawford, Macmillan, and Vignoles 2017), health (Pampel, Krueger, and Denney 2010) and subsequent labour market outcomes (Currie 2009). This has been accompanied by a sustained public policy interest in ‘improving social mobility’ and reducing socio-economic health and academic achievement gaps (Social Mobility Commission 2019). Such bold commitments have been made by policymakers across the western world, including England – the empirical setting for this paper.

In many countries, administrative data – public records about individuals held by government that were not originally collected for research purposes – are being increasingly utilised to understand and address socio-economic differences across a range of outcomes (Connelly et al. 2016; Pattaro, Bailey, and Dibben 2020). A wide body of research in England has used the National Pupil Database (NPD) to add important

CONTACT John Jerrim  J.Jerrim@ucl.ac.uk  Department of Social Science, UCL Institute of Education, University College London, 20 Bedford Way, London, WC1H 0AL, United Kingdom of Great Britain and Northern Ireland

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/02671522.2023.2271930>

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

new evidence on when and how socio-economic differences emerge and how they change throughout childhood (Asaria, Doran, and Cookson 2016; Gorard and Siddiqui 2019; Hire et al. 2018). Such resources are also increasingly being used in policy and practice, including allocating funding to schools (Education and Skills Funding Agency 2018) and making ‘contextual offers’ (i.e. lowering the grades needed) to certain students applying to university (Gorard et al. 2019).

However, one of the limitations with administrative data in studying (and acting upon) socio-economic inequalities is the limited quality of the socio-economic status measures available (Samson et al. 2017; Taylor 2018). Ideally, administrative data would contain at least one of the three most used indicators of socio-economic position (Galobardes, Lynch, and Davey-Smith 2007):

- Social class (parental occupation)
- Parental education
- (Permanent) family-income

Yet as such information is not available in most administrative databases, proxy measures must be used instead. One of the most widely used in education research in England is eligibility for Free School Meals (FSM) – a binary income-based measure used to identify families with low-incomes. Yet a raft of area-based measures – such as the Index of Multiple Deprivation (IMD) – are available as well. Indeed, some have even combined individual and area-level measures to produce a socio-economic status scale (Chowdry et al. 2013). Yet relatively little is known about this collection of measures, including how they compare against one another. This paper hence considers how well a wide array of measures that can be constructed within key administrative databases in England proxy permanent family income.

There has been some previous work investigating the properties of socio-economic proxy measures available in administrative data, though these have typically considered a single indicator in isolation. Hobbs and Vignoles (2010) investigated whether eligibility for FSM in England is a good proxy for low family income. They found FSM children were more likely to be from low-income households, though with some degree of inaccuracy. Taylor (2018) conducted a similar investigation for FSM as a measure of low-income in Wales. He found FSM eligible to be a good proxy, but also that there is ‘*small but significant group of children who could be described as socio-economically disadvantaged and who have low levels of attainment but, for whatever reasons, are not recorded as eFSM [ever FSM eligible]*’. Ilie, Sullivan and Vignoles (2017) investigate the properties of FSM and a set of area-based proxies. They argue that neighbourhood-based measures are not as good as predicting educational achievement as FSM eligibility. Gorard et al (2019) argues that the number of years a child has been eligible for FSM during their time at school is one of the most suitable measures in contextual offers made by universities. On the other hand, they argue neighbourhood measures are less suitable, falling into the problem of the ‘ecological fallacy’ (i.e. a person living in a ‘disadvantaged’ area may not be disadvantaged themselves). Having created a socio-economic index out of a combination of individual and neighbourhood proxies, Campbell et al (2019) report – in an appendix – that this correlates reasonably well with a widely used measure of occupational social class. Sheringham et al (2009) argue that combining information from across two

neighbourhood-based measures (Acorn and IMD) helps to overcome challenges in monitoring health inequalities (at least in the context of sexual health service use). The properties of the Townsend Deprivation Index were investigated by Adams, Ryan and White (2004), who found it to be strongly correlated with individual-level deprivation. Bryere et al (2017) investigate the properties of seven neighbourhood-based proxies in France, including measures equivalent to the Townsend and Carstairs indices widely used in health research in the UK. They argue that such neighbourhood indices are ‘quite good proxies’ for individual deprivation, but that they are also more efficient at measuring individual income than education or occupation, and are more suitable for capturing deprivation than affluence. A different approach was taken by Soobader et al 2001, who investigated whether neighbourhood proxies are less biased measures of individual-level socio-economic status if the unit of geographic aggregation is smaller. They conclude that ‘researchers should be cautious about use of proxy measurement of individual SES even if proxies are calculated from small geographic units’. Similar caution was advised by Link-Gelles et al. (2016).

This paper contributes to this literature in multiple ways. First, the properties of a wide array of possible proxy measures for permanent family income are considered simultaneously. This is important as researchers, practitioners and policymakers will often have a choice of possible proxies, but with little empirical evidence to guide their decision of which to use. Second, we explicitly document the bias within each measure if it is used as a proxy for permanent family income, helping users of such indicators better understand their strengths and limitations. Third, most existing work has investigated how various proxies correlate with income measured at a single point in time. Yet this ignores a wide-ranging economic literature highlighting how it is *permanent* income that is likely to matter (Link-Gelles et al. Jäntti and Jenkins 2015), with measurement error in income data from a single year likely to be severe (Blanden, Gregg, and Macmillan 2013). Finally, we also investigate whether averaging each proxy measure across several years of data provides as better indicator of family income than basing the proxy on a single year of data alone.

2. Family income and its measurement

2.1 Why focus on permanent family income?

This paper focuses on proxies for permanent family income – including permanently *low* income and permanently *high* income. We do so for several reasons. First, income has long been of interest to both economists (Atkinson, Maynard, and Trinder 1983) and sociologists (Breen, Mood, and Jonsson 2015). Second, family income plays a central role in theoretical models of intergenerational persistence and how inequalities are reproduced (Leibowitz 1977). Third, the concept of income is widely understood – and discussed – amongst policymakers. Fourth, interventions are often target specific income groups (Bull et al. 2014; La Valle et al. 2014).

Finally – and perhaps most importantly – we seek to help researchers build on previous work using England’s administrative data. As noted by Connelly, Gayle, and Lambert (2016), researchers should ideally use existing measures that have agreed and well-documented standards. For the NPD in England, the socio-economic variable that has gone through most validation is FSM eligibility. Hobbs

and Vignoles (2010) explored its properties as a proxy for low income in detail, with Taylor (2018) conducting a similar exercise on this ‘*largely income-related measure*’ in Wales. Unfortunately, as we will show, FSM does not capture the middle and top end of the income distribution well – it has only been validated as a measure of low-income. This has consequently led some to develop their own ad-hoc measures when exploring outcomes for higher-income groups (e.g. Chowdry et al 2013; Burgess, Crawford, and Macmillan 2018; Campbell et al. 2019). This paper thus seeks to establish what is best proxy available in administrative data when one is interested in comparing educational outcomes across children from low- and high-income backgrounds, which can then be used to complement analyses using FSM eligibility.

2.2 *The limitations with family income*

Income is just one marker of socio-economic position. Other dimensions of socio-economic status may be equally or more important determinants of educational outcomes; ideally administrative datasets would contain proxy measures capturing these aspects as well. For instance, even when focusing on financial resources, expenditure (both how much and what it is spent on) may be more important for educational outcomes than income per se. Parental social class – as captured by parental occupation – is likely to better capture a family’s social capital and networks than household income, with such factors also an important influence on young people’s educational attainment (Dika and Singh 2002). Likewise, parental education may better capture parental views about education, their cultural capital and capacity to help their children learn. Other factors such as language, cultural norms and school effectiveness – all of which are only partially related to family income – will all also play a role. Thus, although we may find that the measures we consider do not act as good proxies for family income, we cannot rule out the possibility that they may capture some other dimensions of socio-economic position reasonably well. To partially recognise this point, Appendix B provides alternative estimates where we investigate the relationship between each proxy and a multi-dimensional measure of socio-economic status (combining information on parental education, social class and permanent family income). Interestingly, the general pattern of results is qualitatively similar to those presented for permanent family income.

2.3 *Measurement of low-income on this paper.*

Following a long tradition in economic research, our primary interest is in ‘permanent’ (long-run average) rather than ‘transitory’ (single point-in-time) family income. In other words, we are interested in long-run access to financial resources, and particularly the groups who are consistently concentrated at the top and the bottom of the family-income distribution. The empirical analysis therefore considers how well each measure proxies:

- Permanently low-income. Defined as the bottom permanent income quintile.
- Permanently high-income. Defined as the top permanent family-income quintile.
- Permanent income as a continuous variable.

3. Data

3.1 Background

The Millennium Cohort Study (MCS) is a nationally representative longitudinal study. A stratified, clustered survey design was used, with geographic areas (electoral wards) selected as the primary sampling unit, and households with newly born children randomly selected from within (see Plewis 2004 for details). We use data collected between 2000/01 and 2015, when children were nine months, 3, 5, 7, 11 and 14-years-old. In total 19,243 cohort members participated in the first survey (12,224 in England). We focus on the sample for England linked to the NPD due to many of the proxies under investigation being country-specific (e.g. several area-based measures are not designed to be comparable across England, Wales, Northern Ireland and Scotland). The final analytic sample totals 7,439. The MCS wave 6 (age 14) weight is applied whenever possible.

3.2 Measurement of parental income

Across the MCS sweeps, children's parents have been asked several questions about their income. This includes income from work and other sources (e.g. benefits, investments) with information recorded for both mothers and fathers. Although the income data are self-reported – and unlikely to be free from error (Moore, Stinson, and Welniak 2000) – best practice has been followed in their collection. The survey organisers have used the information provided by respondents to harmonise the data across sweeps as far as possible. There is also a degree of non-response to the questions asked about income. For instance, Hawkes and Plewis (2008) note that *'about 6% of main respondents and 6% of partners did not provide income data at sweep one'*, finding that this seems to be related to how individuals earn their living, who they live with, geographic location and ethnicity. Non-response to questions about income may reduce the representativeness of the sample. To derive permanent income we average total household income when the child was age nine months, 3, 5, 7, 11 and 14. Previous research has suggested this is a sufficient time horizon to provide a good measure of permanent income (Gregg, Macmillan, and Vittori 2017).

3.3 Data linkage

Access to a secure version of the MCS data was made available by the Centre for Longitudinal Studies Data Access Committee. This included information on full postcodes across MCS sweeps. Such detailed information was necessary so that the area-based proxies could be derived at the most fine-grained geographic level possible. Parents were also asked for their consent for their children's records to be matched to the NPD. In England, the overall rate of consent was 94%, with 89% of these records then successfully linked (Breen, Mood, and Jonsson 2015). Although these consent and linkage rates are high, the fact they are not perfect may reduce the representativeness of the sample used in the analysis. Nevertheless, for those that consented, our data includes information on FSM eligibility for each year the child was registered at school and their GCSE grades.

4. The proxy measures

Table 1 provides an overview of the proxy measures we consider. A full description is provided in the online supplementary material (Appendix A) with a short summary provided below. The proxies we consider are measured at one of five levels:

- Individual = Microdata relating to the individual child or their parents.
- Postcode = Geographic areas that have a median size of 13 households and 31 residents (Acorn Technical Guide 2017).
- Output Area = Geographic areas that includes around 300 individuals in approximately 125 households.
- Lower Super Output Area = Geographic areas that includes around 1,500 individuals in approximately 650 households.
- Middle Super Output Area = Geographic areas that includes around 7,500 individuals in approximately 3,000 households.

4.1 Acorn

Acorn is a geodemographic classification system that combines information from the Land Registry, administrative and commercial data to divide each postcode in the UK

Table 1. The proxy measures investigated in this paper.

Measure	Abbreviation	Level measured at	Permanent measure	Proxy for low-income		Proxy for high-income	
				Definition	% in group	Definition	% in group
Index of Multiple Deprivation	IMD	LSOA	Mean over time	Bottom quintile	20%	Top quintile	20%
ACORN	Acorn	Postcode	Mode over time	Category 4/5	49%	Category 1/2	23%
Free-school meals	FSM	Individual	-	-	17%	-	83%
Years of free school meals	FSM	Individual	Mean over time	% time FSM eligible	20%	Never FSM	67%
Income Deprivation Affecting Children	IDACI	LSOA	Mean over time	Bottom quintile	20%	Top quintile	20%
Carstairs Index	Carstairs	LSOA	Mean over time	Bottom quintile	20%	Top quintile	20%
Output Area Classification	OAC	LSOA	Mode over time	Groups 7,8 3a-3c, 4b	38%	See Appendix A	18%
IFS index	IFS	Individual/postcode	Mean over time	Bottom quintile	20%	Top quintile	20%
Townsend index	Townsend	OA	Mean over time	Bottom quintile	20%	Top quintile	20%
Young Participation by Area Rate	POLAR	MSOA	Mean over time	Bottom quintile	20%	Top quintile	20%
Tracking underrepresentation by area	TUNDRA	MSOA	Mean over time	Bottom quintile	21%	Top quintile	20%
Transitory income (age 14)	Income (age 14)	Individual	N/A	Bottom quintile	20%	Top quintile	20%

Notes: Individual = Microdata that relate to the individual child or their parents. Postcode = Geographic areas that have a median size of 13 households and 31 residents (Acorn Technical Guide 2017). Output Area (OA) = Geographic areas that includes around 300 individuals in approximately 125 households. Lower Super Output Area (LSOA) = Geographic areas that includes around 1,500 individuals in approximately 650 households. Middle Super Output Area (MSOA) = Geographic areas that includes around 7,500 individuals in approximately 3,000 households.

into 62 Acorn types. These are based on information such as house sales, rentals, social housing, information about residents, benefits claimants, census and lifestyle data (Acorn Technical Guide 2017). We follow the University of Oxford outreach team (University of Oxford 2022), with Acorn Category 4 and 5 used as a proxy for low-income (49% of the population) and Acorn Category 1 and 2 used as a proxy for high-income (23% of the population).

4.2 Output Area Classification (OAC)

The OAC is a geodemographic classification system developed by the Office of National Statistics. The data are open source, with each census Output Area classified into one of eight OAC groups, 26 groups and 76 sub-groups. The OAC groupings are categorical – not ordinal – and based on the demographic structure, household composition, housing type, socio-economic and employment situation of the area.

We follow the University of Cambridge (2019) outreach team to create our OAC proxy for low-income, which captures 18% of the population. To proxy high-income, we use the MCS to establish the OAC sub-groups with the highest average weekly pay, and then take those that together account for the top 20% of the permanent family income distribution. See Appendix A for further details.

4.3 Index of Multiple Deprivation (IMD)

The IMD is the official measure of relative deprivation used in England. It is comprised of seven deprivation domains, measured at Lower Super Output Area level, that are combined (with unequal weight) to form the final scale. In total, the IMD combines information from across 39 separate indicators, sourced from government administrative data and the census (see Ministry of Housing 2019). We use the bottom/top IMD quintiles to proxy low/high income.

4.4 Income Deprivation Affecting Children (IDACI)

IDACI is a sub-scale of the income domain of the IMD. It ranks the proportion of 0–15-year-old children living in income deprived families in each Lower Super Output Area. See Ministry of Housing (2019) for further details. We use the bottom/top IDACI quintiles to proxy low/high income.

4.5 Eligibility for Free School Meals (FSM)

FSM are a means-tested benefit, though the criteria used to determine eligibility has changed over time (Hobbs and Vignoles 2010). Children are flagged as ‘eligible’ for FSM in the NPD only if they are both eligible for and claiming FSM (Hobbs and Vignoles 2010). It is possible to calculate the proportion of time children have been FSM eligible throughout their time at school. FSM is the only proxy we consider that is solely based on individual level data.

When using a single year of data, low-income is proxied by those eligible for FSM in that year (17% of the population), with high-income proxied by those who are not (83%).

To create a ‘permanent’ FSM measure, we calculate the proportion of time at school (between ages 5 and 16) children were FSM eligible. Low-income is then proxied as the 20% of children who were eligible for FSM for the greatest proportion of their time at school, with high-income proxied by the 67% of children who were never FSM eligible.

4.6 Participation of local areas (POLAR)

POLAR is an indicator of university participation by local area – capturing how likely young people are to participate in higher education depending on where they live. The ‘youth participation rate’ is first calculated as the number of 18/19-year-olds from a given area who enter higher education, which is then divided by the total number of 18/19-year-olds living in that area. This continuous measure is then divided into five quintiles that form the POLAR groups.

It should be noted that POLAR was initially developed to represent a particular socio-economic structure (higher education participation) at the area level. Although it has sometimes been co-opted as proxies for individual level disadvantage, there is a detailed literature criticising such use, particularly in the allocation of bursaries and widening participation initiatives (e.g. Boliver, Gorard, and Siddiqui 2021). This has led the regulator in the UK (the Office for Students) to caution universities against using it without other markers for individual disadvantage (Office for Students 2019). Thus, although we do not anticipate POLAR to correlate highly with family income, it is worthwhile demonstrating this empirically given how this measure has sometimes been used.

4.7 Tracking underrepresentation by area (TUNDRA)

TUNDRA is an indicator of university participation by local area developed by the Office for Students (2021). Specifically, it is the proportion of 16-year-olds from state schools who enter higher education divided by the total number of 16-year-olds within a given area – producing a continuous measure that is then divided into five TUNDRA quintiles. As with POLAR, we do not anticipate this measure to correlate particularly well with family income, given how it has been designed. Nevertheless, as it has sometimes been used to make important individual-level decisions, we believe its capacity to proxy family income is still worth demonstrating. We thus investigate the properties of the bottom/top TUNDRA quintile as a proxy for low/high income.

4.8 Carstairs index

The Carstairs index (Carstairs and Morris 1991) combines four census variables recorded at Lower Super Output Area level: (a) male unemployment; (b) lack of car ownership; (c) overcrowding and (d) low social class. Each of these variables is standardised to mean zero and standard deviation one, and then summed together to create the final scale. Data for the Carstairs index are available from Wheeler (2019). Our analysis investigates the properties of the bottom/top quintile of the Carstairs index as a proxy for low/high income.

4.9 Townsend index

The Townsend index (Townsend, Phillimore and Beattie 1988) combines four census variables recorded at the Output Area level: (a) unemployment; (b) lack of car ownership; (c) overcrowding and (d) non-home ownership. The unemployment and overcrowding indicators are first log-transformed, with each of the four variables then standardised to mean zero and standard deviation one. These standardised scores are then summed to create the final scale. See Yousaf and Bonsall (2017) for further details. Our analysis focuses on the bottom/top quintiles as a proxy for low/high income.

4.10 The Institute for Fiscal Studies (IFS) index

The IFS measure was first developed by Chowdry et al (2013) and since been used in a handful of papers (e.g. Burgess, Crawford, and Macmillan 2018; Campbell et al. 2019). It combines the following information via a principal components analysis (PCA) with the first component used to construct the scale:

- Eligibility for FSM at age 16.
- IMD (measured at Lower-Super Output Area level).
- Acorn type (measured at postcode level).
- Neighbourhood socio-economic status, education level and housing tenure (measured at the Output Area level).

Unfortunately, Chowdry et al (2013) provide few other details about how the scale is constructed (e.g. percent of variance explained, values of loadings) although they do demonstrate its associations with various markers of socio-economic background.¹ Likewise, other studies that have used this measure report only limited details (Burgess, Crawford, and Macmillan 2018; Campbell et al. 2019). One challenge is that principal components analysis is a model-based approach that explores the association between manifest (observed) variables to construct a latent scale – and is hence based upon a particular dataset at a given point in time. If the scale is then re-created using a different dataset covering a different timeframe, there are likely to be small changes in the associations between the manifest variables which will produce slightly different estimates.

Given that we cannot directly reproduce the scale others have used from the information published, we proceed by estimating new polychoric PCA models using the MCS data, including the same variables used by others as listed above.² The first component explains 62% of the variance, with loadings for the age 14 wave reported in Table 2 (see Appendix E for the loadings from all waves). We then standardise the scale to mean zero and standard deviation one. We investigate the bottom/top quintiles as potential proxies for low/high income.

Table 2. Details about the polychoric PCA model used to construct the IFS scale (wave 6/age 14).

Component	Eigenvalues	% explained
(a) Eigenvalues		
1	3.70	62%
2	0.89	15%
3	0.65	11%
4	0.36	6%
5	0.24	4%
6	0.16	3%
Variable	Loading	
(b) Loadings		
Not FSM	0.168	
FSM	-0.357	
IMD	-0.459	
Acorn	-0.432	
Neighbourhood socio-economic status	0.448	
Neighbourhood education level	0.364	
Neighbourhood homeownership	0.408	

Notes: See Appendix A for eigenvalues and loadings from the other MCS waves and the code used to construct the scale.

5. Methodology

5.1 Estimating the correlation between each proxy and permanent family income

One criterion for selecting a good proxy is that it should be strongly correlated with the construct of interest (Lewis-Beck, Bryman and Liao 2004). For each proxy we therefore estimate how strongly it is associated with low-income, high-income and a continuous measure of permanent income. Where the two measures are both continuous (e.g. when comparing permanent income to the IMD) Pearson correlation coefficients are presented. Point biserial correlations are estimated when one variable is categorical and the other continuous (e.g. comparing FSM to permanent income) while polychoric correlations are used when both measures are categorical (e.g. when comparing FSM to low income). We estimate these correlations using both a ‘transitory’ measure of the proxy (taken when the cohort member was age 14³) and a ‘permanent’ measure (an average across six time points spanning 13 years). This will reveal whether better proxy measures of permanent income can be derived if data is available over time.

To further explore the association between the proxy measures and low/high income, we convert each proxy into binary form. This is done using established thresholds, or simply taking the top/bottom quintile of the distribution, as summarised in Table 1. We then calculate the ‘true-positive’ rate (e.g. the percent of cohort members the proxy measure correctly identifies as low-income) and the ‘false-positive’ rate (e.g. the percent of cohort members the proxy measure incorrectly identifies as low-income when they are not) for each indicator in turn.

A summary of this approach is provided for one indicator (the IMD). To begin, the IMD is divided into quintiles, with the top fifth of the distribution taken as a proxy for high income and the bottom fifth as a proxy for low income. These are then cross tabulated with our ‘true’ measure of low-income as illustrated in Table 3. Figures in the top-left cell (shaded in green) provide the ‘true positive’ rate – the percent of children in

Table 3. An illustration of how the true-positive and false-positive rate for low-income is calculated using the index of Multiple Deprivation (IMD).

	Low income	Not low income
Bottom IMD quartile	43.7%	14.1%
Not in the bottom IMD quartile	56.3%	85.9%
	100%	100%

Notes: True-positive rate highlighted in the top-left hand corner (shaded green). The false-positive rate is in the top-right hand corner (highlighted in red).

low-income households that the proxy (bottom IMD quintile) correctly identifies. Conversely, figures in the top-right cell (shaded in red) provides the ‘false positive’ rate; the percent of cohort members the IMD proxy incorrectly identifies as from a low-income background when they are not. With respect to the IMD, Table 3 illustrates that the true-positive rate is greater than the false-positive rate. The ideal proxy will of course maximise the former (true positives) while minimising the latter (false positives). We will compare how the proxies perform in this respect using a scatterplot, known formally as the Receiver Operating Characteristic (ROC) space (Hajian-Tilaki 2013). This will be explained in further detail when presenting the results in the following section.

One limitation is that the percent of true-positive and false-positive cases for a proxy depends on where one chooses to ‘cut’ the data (e.g. the point along the IMD distribution one should pick for it to be the best proxy for low income). We therefore also estimate the true-positive and false-positive rates for each proxy when the ‘optimal’ cut-point is used (with ‘optimal’ meaning maximising the true-positive rate and minimising the ‘false-positive’ rate).⁴ This is implemented via the Stata package *cutpt* (Clayton 2013).⁵ Appendix D provides information on the percentage of the population falling into the low-income and high-income proxy groups when the optimal cut-point is used.

5.2 Investigating bias between key demographic groups

The ideal proxy should also only capture the unobserved variable of interest (permanent family income) and not other characteristics of the individual. If this is not the case, then the proxy will be biased in favour of one group compared to another – at least as a measure of family income. We explore potential bias in each measure as a proxy for permanent family income via the following OLS regression model:

$$Inc_i = \alpha + \beta.Characteristic_i + \gamma.Proxy_i + \varepsilon_i \quad (1)$$

Where:

Inc_i = Permanent family income.

$Characteristic_i$ = One of the background characteristics we explore whether the proxy is biased towards/against.

$Proxy_i$ = The proxy variable in question.

i = child i.

The parameter of interest is β ; this captures the relationship between the characteristic in question (e.g. gender) and permanent family income, once the proxy

measure has been controlled. For the ideal proxy, the estimated β parameter would be zero – after accounting for differences in the proxy measure, there would be no systematic differences in permanent family income between groups. On the other hand, the greater the absolute value of β , the greater the bias in the proxy measure (as a marker for family income). For instance, say that after controlling for the IMD, there continues to be differences in permanent family income by ethnicity. This would indicate that the IMD partially captures the effect of ethnicity, rather than the family income alone.

The model outlined in (1) will be estimated separately for each proxy measure and following background measures:

- Gender
- Ethnicity (White/not white)
- Single parent household
- Geographic location (live in London or not)
- Home ownership (yes/no)
- Young mother (gave birth at age 21 or below)

We are interested in factors such as geographic region and homeownership given that several proxy measures use local area-level characteristics which may not hold the same meaning across different geographic regions. For instance, individuals may be able to afford to live in a more affluent neighbourhood (and hence receiving a higher score on the proxy measure) if they choose to rent rather than buy their own home. We recognise that more fine-grained comparisons would be possible for some of the characteristics considered (e.g. location, ethnicity). This would however substantially increase the number of comparisons made, and thus the size and complexity of results tables. Sample sizes would also become small for some ethnic groups.

The analysis presented in (1) will also be replicated for our binary low/high income measures using a linear probability model, controlling for each proxy measure in turn (also in its binary form). These models will reveal whether certain groups are more likely to be categorised as low/high income by each proxy than would be the case if the true measure of interest (permanent family income in a categorised form) were available.

5.3 Estimating income achievement gaps

A good proxy measure should also have good predictive validity. We operationalise this concept as the extent that each proxy can replicate permanent family income differences in academic achievement. In other words, the academic achievement of different groups (as defined by each proxy) should be similar to those when using permanent family income. We therefore investigate the percent of children from low- and high-income backgrounds who achieved five General Certificate of Secondary Education (GCSE) A*-C grades using each proxy measure, and how this compares to when using permanent family income.

Table 4. The correlation between different proxy measures and permanent family income.

Measure	Type	Correlation with low-income	Correlation with permanent income	Correlation with high-income
IMD	Age 14	0.47	0.44	0.52
	Permanent	0.50	0.48	0.58
FSM	Age 14	0.60	0.33	-
	Permanent	0.69	0.44	-
IDACI	Age 14	0.48	0.44	0.52
	Permanent	0.50	0.49	0.58
YPR/POLAR	Age 14	0.22	0.38	0.47
	Permanent	0.20	0.41	0.51
IFS	Age 14	0.51	0.55	0.63
	Permanent	0.52	0.58	0.67
ACORN	Age 14	0.56	0.54	0.66
	Permanent	0.61	0.59	0.70
Carstairs index	Age 14	0.47	0.46	0.53
	Permanent	0.50	0.49	0.59
Townsend index	Age 14	0.50	0.45	0.47
	Permanent	0.53	0.50	0.53
TUNDRA	Age 14	0.17	0.30	0.38
	Permanent	0.13	0.31	0.41
OAC	Age 14	0.46	0.41	0.55
	Permanent	0.44	0.47	0.40
Income age 14	Age 14	0.73	0.81	0.89

Notes: Shading should be read vertically. Higher correlations are in green shades; lower correlations in red shades.

6. Results

6.1 Correlations

Table 4 presents the correlation between each proxy and permanent income. Shading should be read vertically, with green (red) cells indicating whether the correlation is stronger (weaker). The strength of the correlation is reported when the proxy is measured at age 14 and the ‘permanent’ measure (averaged across all MCS sweeps). The correlation between age 14 income and permanent family income is presented in the bottom row to facilitate interpretation of results (i.e. it represents perhaps the best possible benchmark one could expect a proxy to meet).

There are four key points. First, TUNDRA and POLAR are only weakly correlated with permanent income and (particularly) permanent low-income; the correlation coefficients are around 0.2 to 0.4.

Second, FSM eligibility (particularly the ‘permanent’ measure combining information across several years) has the strongest correlation with permanent low-income across all the measures considered. In fact, the correlation for time-average FSM and permanent low-income ($r = 0.69$) is almost the same as between age 14 (‘transitory’) low-income and permanent low-income ($r = 0.73$). Yet the correlation between FSM and permanent income is notably weaker ($r = 0.44$) and lower than for several of the other proxies. This reflects FSM being a coarse indicator that only provides information about the bottom of the income distribution; it does not discriminate well between low, middle and high-income groups.

Third, there is not a lot to choose between the other measures. The correlation between each proxy and permanent income is generally around 0.45–0.60. Acorn and the IFS measure have slightly stronger correlations than some others (e.g. IMD, Carstairs index, Townsend index) though the difference is relatively small.

Finally, the only proxy for which there is a meaningful difference between the age 14 and time-average (permanent) versions is FSM. This likely reflects its binary nature within a single year, with more information (and variation) gained when taking averaged over time. Otherwise, in terms of the correlation with permanent income, it does not seem to matter if one averages the proxy over time, or if data from a single timepoint is used.

6.2 True-positive and false-positive rates

Figure 1 presents evidence on the ability of each measure to identify young people living in permanent low-income. The vertical axis plots the ‘true-positive’ rate; the percentage of children that the proxy correctly identifies as being from a low-income background. Meanwhile, the false-positive rate is plotted along the horizontal axis; the percent of children not from a low-income background that the proxy mistakenly classifies as so. Ideally, a proxy should maximise the former (true positive) while minimising the latter (false positives) – meaning better performing proxies will tend to sit towards the top-left corner. The dashed 45-degree line is where the true-positive and false-positive rates are equal – where the proxy is essentially of no use in distinguishing low-income from not low-income groups.

Where each proxy falls on this graph is a function of two factors: (a) how well each proxy captures permanent income; (b) for continuous proxy measures, the threshold below which a child is classified as from a low-income background. Two versions are therefore presented. Panel (a) bases the grouping/cut-point for each proxy on how it has often been used in research, policy and/or practice (for specific details see section 4 and Appendix A). In contrast, panel (b) takes an empirical approach to determining the ‘optimum’ cut-point – defined as the point along the proxy measure which maximises the true positives and minimises the false positives.⁶

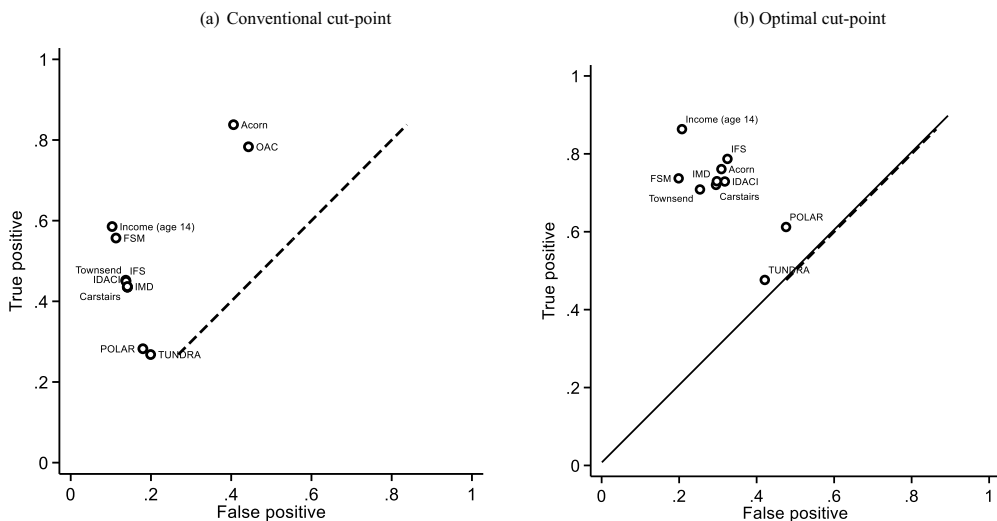


Figure 1. True-positive and false-positive rates for detecting low-income using different proxy measures. Notes: True-positive rate is presented on the vertical axis; this captures the percent of low-income families correctly identified by the proxy. The false-positive rate is plotted along the horizontal axis, capturing the percent of families the proxy identifies as coming from a low-income background when they do not. The 45-degree line illustrates where the true-positive and false-positive rate is equal, meaning the proxy is of no use in identifying low-income groups. The ideal proxy would sit in the top-left corner of the graph.

The first notable feature of [Figure 1](#) is that it reaffirms the problems with using POLAR and TUNDRA as proxies for low-income. Both sit close to the 45-degree line, with the true positive rate almost identical to the false positive rate. Clearly, these two measures are not good at identifying young people from low-income backgrounds, regardless of where the ‘cut-point’ is drawn.

The Acorn and OAC proxies stand out in panel (a) for a different reason. Compared to all other indicators, both the true positive and false positive rates are a lot higher. For instance, around 80% of children from low-income homes will be identified as such using Acorn – at least when following how this measure has been used by the University of Oxford (2022). Yet the false positive rate – not low-income children that Acorn classifies as low-income – also stands at around 40% (double that of most other measures). This is a function of how the low-income group using Acorn has been defined; it encompasses half (49%) of the population.⁷ As can be seen in panel (b), the true positive and false positive rate for Acorn is very similar to other proxies when the ‘optimal’ cut-point is used. This highlights a key point when using proxy measures to identify particular income groups. It is not only important to consider how well the proxy captures the underlying construct of interest (e.g. permanently low family income) but also the ‘cut-point’ selected.

Finally, there is again not a great deal to choose between the other proxy measures. FSM eligibility has a higher true positive rate than the other proxies (and a similar false positive rate) when conventional cut-points are used (see panel a). This advantage disappears in panel (b), however, when the optimal cut-point is selected instead. Hence the advantage of FSM in identifying children from low-income backgrounds over other measures seems to largely be a function of where the conventional ‘cut-point’ on the other proxies has been drawn. Indeed, in panel (b) – with the exception of POLAR and TUNDRA – there is no clear case for preferring any one of the proxies over another.

Analogous results for identifying children from high-income backgrounds can be found in [Figure 1](#). Unsurprisingly, the time a child has been FSM eligible stands out from other data points. It has a true-positive rate of almost one (due to almost all children from high-income backgrounds having never been eligible for FSM) but also a much higher false-positive rate (due to many children who have never been FSM eligible not coming from particularly high-income backgrounds). Hence FSM is an outlier due to its coarse nature, with it not being well-suited to capturing the middle or top of the income distribution.

For the remaining proxies, using conventional cut-points in panel (a), there is little difference in the false positive rate for income-affluence; for each this stands at quite a low-level (between 0.1 and 0.2). However, there is more variation in terms of the true positive rate, with Acorn (0.62) and the IFS measure (0.54) notably higher than for other indicators such as TUNDRA (0.37), OAC (0.39), POLAR (0.44) or the Townsend Index (0.45). On the other hand, in panel (b) where the optimal cut-point is used, both the true positive and false positive rates tend to be somewhat higher. Although most of the measures now have a similar true positive rate when the optimal cut-point is used (except POLAR and TUNDRA which are somewhat lower, and FSM which is somewhat higher) the IFS and Acorn measures now have a slightly lower false positive rate than most others. Together, this suggests that the Acorn and IFS measures may have some slight advantage at proxying high-income than most of the local area-based alternatives (e.g. IDACI, IMD, Townsend, Carstairs), and major advantages relative to POLAR, TUNDRA and time-averaged FSM.

6.3 Investigation of bias

Results considering the bias in each proxy as a measure of permanent low-income can be found in Table 5. Analogous results for permanent high-income can be found in Table 6, and those for a continuous measure of permanent income in Appendix C. For instance, the figures in Table 5 refer to how much more likely the group in question (e.g. single versus two-parent households) are to actually live in permanent low-income, having controlled for differences between groups in the proxy measure. Ideally, figures would be close to zero, indicating the proxy captures all differences in low-income between groups. The shading of cells should be read vertically, with green cells indicating less bias in the proxy (i.e. values closer to zero) as an indicator of family income.

For many of the characteristics considered, the absolute bias in each proxy is quite large. In other words, on this criterion, none of the proxies perform particularly well. Take the results for single parent households, for example. We find this group to generally be around 20% points more likely to be low-income than two-parent families, even after accounting for differences between these groups on the proxy measures. This reveals how the proxies are not a ‘pure’ measure of permanent family income; they partly capture some other aspects about individual circumstances, such as ethnicity, housing tenure, family structure and (parental) age. Whether this matters is likely to depend on the context the proxy is being used.

Nevertheless, out of all the proxies considered, permanent (time-averaged) FSM eligibility seems to outperform the others – at least in terms of having minimal bias on the six demographic characteristics considered. The clearest examples are for family structure (single versus two parent families) and housing tenure (renter versus homeowner) where the bias in FSM as an indicator of long-term low-income is much lower

Table 5. Bias in the proxies as a measure of permanent low-income.

Measure	Type	London	Ethnic minority	Single parent	Renter	Male	Young mother
IMD	Age 14	10%	16%	22%	26%	-1%	25%
	Permanent	11%	15%	21%	26%	-2%	24%
FSM	Age 14	9%	17%	16%	21%	-1%	23%
	Permanent	5%	15%	12%	17%	-1%	20%
IDACI	Age 14	6%	16%	22%	26%	-1%	25%
	Permanent	6%	16%	21%	25%	-1%	24%
YPR/POLAR	Age 14	14%	23%	24%	30%	-1%	28%
	Permanent	13%	23%	24%	30%	-2%	29%
IFS	Age 14	9%	17%	21%	25%	-1%	24%
	Permanent	10%	17%	21%	25%	-1%	23%
ACORN	Age 14	11%	17%	19%	24%	-1%	22%
	Permanent	11%	16%	19%	23%	-1%	20%
Carstairs index	Age 14	5%	13%	22%	26%	-1%	25%
	Permanent	5%	11%	22%	26%	-2%	25%
Townsend index	Age 14	2%	12%	21%	25%	-1%	25%
	Permanent	1%	10%	21%	25%	-1%	24%
TUNDRA	Age 14	13%	23%	24%	30%	-1%	29%
	Permanent	13%	24%	25%	30%	-3%	26%
OAC	Age 14	11%	17%	22%	27%	-1%	25%
	Permanent	11%	16%	21%	25%	-1%	23%
Income age 14	Age 14	8%	15%	18%	24%	-2%	24%

Notes: Figures indicate how much more likely the group is to have permanently low-income, conditional on the proxy measure. For instance, Londoners are around 14% points more likely to actually have low income than those living elsewhere in England, conditional upon age 14 POLAR as a proxy. Values close to zero indicate less bias in the proxy measure and are shaded in green (red shading is where the bias is greater).

Table 6. Bias in the proxies as a measure of permanent high-income.

Measure	Type	London	Ethnic minority	Single parent	Renter	Male	Young mother	Young father
IMD	Age 14	7%	-7%	-18%	-25%	2%	-16%	-14%
	Permanent	6%	-5%	-17%	-23%	1%	-14%	-12%
FSM	Age 14	6%	-8%	-16%	-26%	1%	-16%	-14%
	Permanent	8%	-5%	-12%	-21%	1%	-11%	-13%
IDACI	Age 14	8%	-6%	-18%	-25%	2%	-16%	-14%
	Permanent	8%	-4%	-17%	-23%	1%	-14%	-12%
YPR/POLAR	Age 14	0%	-12%	-20%	-27%	1%	-17%	-15%
	Permanent	-3%	-12%	-20%	-26%	1%	-16%	-14%
IFS	Age 14	6%	-7%	-17%	-23%	1%	-14%	-12%
	Permanent	6%	-5%	-16%	-21%	1%	-12%	-11%
ACORN	Age 14	3%	-7%	-15%	-21%	1%	-13%	-11%
	Permanent	4%	-5%	-15%	-19%	1%	-11%	-9%
Carstairs index	Age 14	9%	-6%	-18%	-25%	1%	-16%	-14%
	Permanent	9%	-4%	-17%	-23%	1%	-14%	-12%
Townsend index	Age 14	9%	-6%	-18%	-26%	1%	-17%	-15%
	Permanent	9%	-5%	-18%	-24%	2%	-15%	-13%
TUNDRA	Age 14	-3%	-14%	-20%	-28%	1%	-19%	-16%
	Permanent	-6%	-17%	-20%	-28%	1%	-18%	-16%
OAC	Age 14	4%	-8%	-18%	-26%	1%	-16%	-14%
	Permanent	3%	-7%	-18%	-25%	1%	-15%	-13%
Income age 14	Age 14	2%	-4%	-4%	-13%	1%	-9%	-9%

Notes: Figures indicate how much more likely the group is to have permanently high-income, conditional upon the proxy measure. Values close to zero indicate less bias in the proxy measure and are shaded in green (red shading is where the bias is greater).

than for the other measures. Moreover, for none of the six characteristics examined does permanent FSM perform poorly as an indicator of low-income relative to the other proxies available. Hence, in terms of selecting a single indicator of long-term low-income that has the least bias against key demographic groups, eligibility for FSM is likely the best pick.

Otherwise, there are two final features to note. First, two proxies that are widely used in medical research appear to be subject to less bias against Londoners and ethnic minorities than the other alternatives – the Carstairs index and the Townsend index. Second, the POLAR and TUNDRA measures once again do not proxy permanent income well – standing out as having big differences between demographic groups when used to proxy family income.

6.4 Estimation of achievement gaps

Figure 2 investigates the magnitude of income-achievement gaps using each proxy. The percent of teenagers from low-income backgrounds who achieved five A*-C GCSE grades (according to each proxy) is plotted along the horizontal axis, with analogous figures for high-income backgrounds on the vertical axis. The square marker labelled ‘income’ presents the results for permanent low/high-income groups (i.e. the ‘true’ income-achievement gap that we want the proxies to replicate). As such, proxies that fall closer to the ‘income’ data point more closely replicate the desired results.

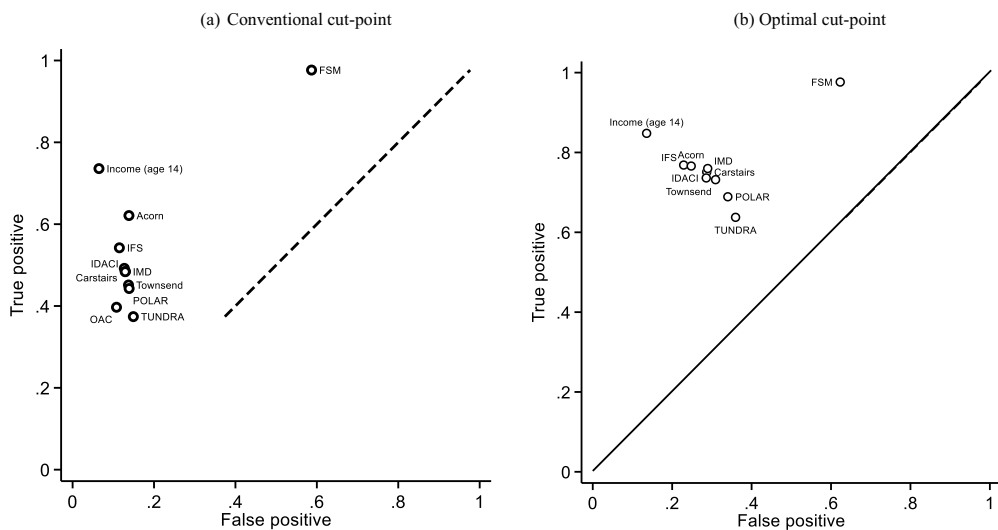


Figure 2. True-positive and false-positive rates for detecting high income using different proxy measures.

Notes: True-positive rate is presented on the vertical axis; this captures the percent of high-income families correctly identified by the proxy. The false-positive rate is plotted along the horizontal axis, capturing the percent of families the proxy identifies as coming from a high-income background when they do not. The 45-degree line illustrates where the true-positive and false-positive rate is equal, meaning the proxy is of no use in identifying high-income groups. The ideal proxy would sit in the top-left corner of the graph.

It is immediately notable how the results for FSM do not match those using permanent family income; the achievement of both high and low-income groups are underestimated. This partly stems from the coarseness of FSM and its lack of flexibility; it does not discriminate well between middle and high-income groups, which limits its attractiveness as a measure to understand advantages enjoyed by those from high-income backgrounds. This is an important – and often underappreciated – limitation of FSM.

A second key feature of [Figure 3](#) is that, when using area-based proxies, the family income-academic achievement gap is attenuated. Specifically, the percent of low-income teenagers getting good school grades is overestimated, while for high-income teenagers it is underestimated. Take the results for the age 14 IMD, for example. Using this proxy, 37% of teenagers from low-income backgrounds achieve five A*-C grades, compared to 70% of those from high-income backgrounds (an income-achievement gap of 33% points). This is compared to a 41% point income-achievement gap when permanent income is used. A similar discrepancy – if not larger – can also be observed for the other area-based measures. In other words, use of area-based proxy measures can lead researchers and policymakers to underestimate the magnitude of academic achievement gaps between children from high- and low-income backgrounds.

Interestingly, this problem of attenuation seems to be greater in panel a of [Figure 3](#) (using the age 14 measure of the proxy) than panel b (permanent, time-average of the proxy). In particular, most of the proxies move northwards in panel

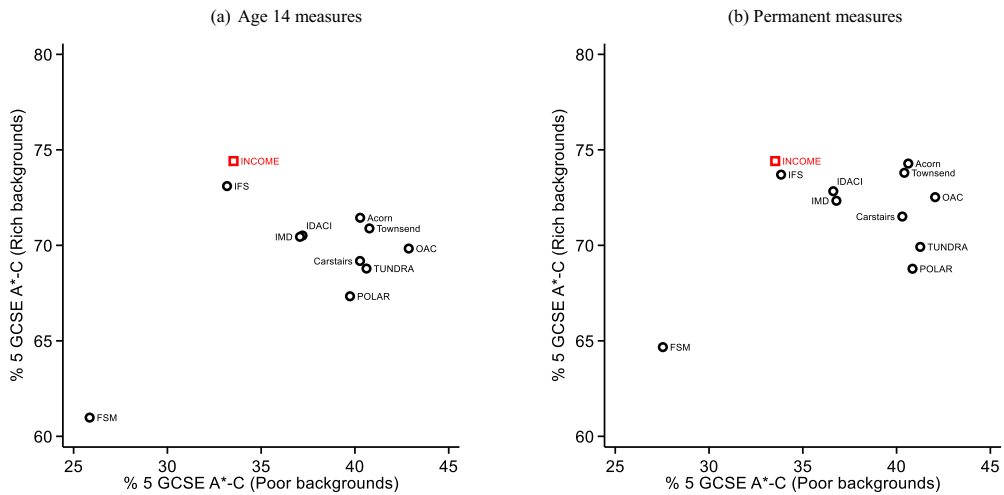


Figure 3. How well do the proxy measures capture permanent family income gaps (top versus bottom quintile) in academic. Notes: Figures refer to the percent of children who achieve five A*-C GCSEs. Results for each proxy using the ‘conventional’ cut-point described in Table 1 and Appendix A. Results for permanent-income illustrated using a red square.

b compared to panel a, with the ‘permanent’ measures leading to higher (and more accurate) estimates of the academic achievement of high-income groups. This suggests that, when it comes to estimating differences in outcomes between high and low income groups, there may be some benefit to researchers having access to proxies derived from address histories over a period of 10–15 years.

However, perhaps the standout feature of Figure 4 is the similarity of results using the IFS proxy to those using permanent income. In both panel (a) and (b) the IFS and permanent income data points sit closely together; much more so than the other proxies. (This is particularly the case in panel a when just the age 14 versions of the proxies are used). This suggests that the IFS measure – which combines information on FSM eligibility, Acorn type and local area census data – may be particularly useful to researchers looking to estimate differences in academic achievement between children from different income backgrounds.

This point is reiterated by Figure 4. The horizontal axis plots percentiles of the permanent income/IFS proxy distribution, while the vertical axis presents the percent of children achieving five A*-C GCSE grades. The black (grey) lines hence illustrate how academic achievement varies for young people whose families sit at different points of the permanent income (IFS proxy) distribution. Importantly, the two lines track each other very closely; the IFS measure produces results that are very similar to those using permanent family income across the distribution. This suggests that, if researchers and/or policymakers are interested in estimating differences in outcomes between low, middle and high-income earners, the IFS proxy is likely to be a valuable resource.

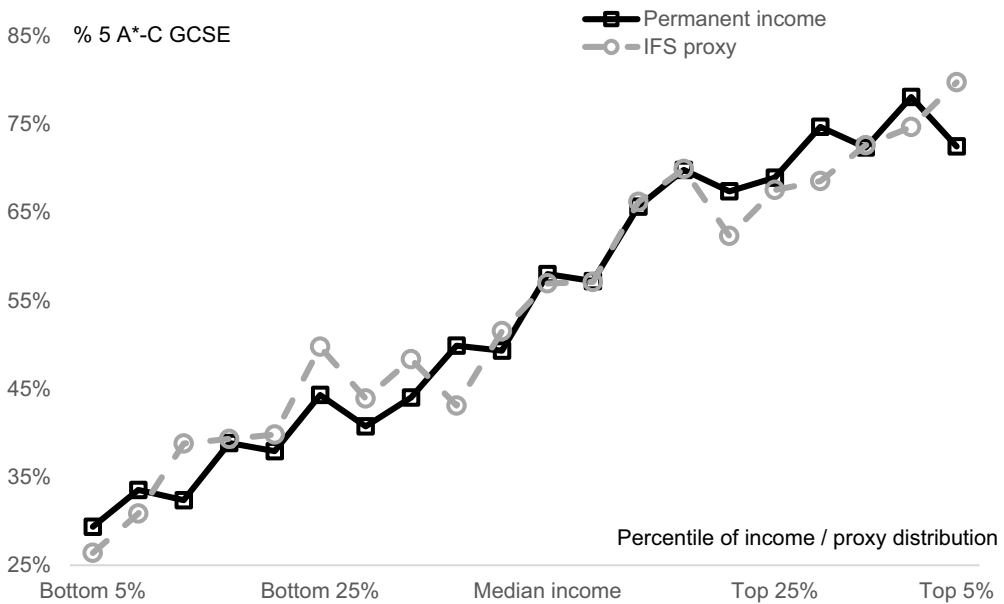


Figure 4. The percent of children achieving five A*-C grades by position in the permanent income distribution.

Notes: The horizontal axis plots the position in the permanent income (or permanent IFS proxy) distribution. Figures on the vertical axis indicate the percent of children achieving five A*-C GCSEs.

7. Conclusions and recommendations

Administrative databases are widely used to inform public policy and in academic research (Mohammed and Stevens 2007). Such resources are being increasingly used to understand inequalities in education and health outcomes (Asaria, Doran, and Cookson 2016), including access to health services (Charlton et al. 2013) academic achievement (Gorard and Siddiqui 2019) and access to university (Chowdry et al. 2013). One major limitation, however, is that such resources contain limited information about socio-economic background. Proxy measures are hence typically used instead.

Yet, despite their widespread use, relatively little is known about how well these proxies capture key aspects of socio-economic position such as family income. There is a particular dearth of information in how they compare to one another in this respect. This paper has added this evidence to the literature, investigating the properties of a set of possible proxy measures for permanent family income that could easily be made available within administrative data.

We find that the number of years eligible for FSM is the best available marker for a child coming from a low-income background. Yet, as this proxy is focused upon the bottom end of the income distribution, it is not well-suited to capturing differences in outcomes between low, average and high-income groups. For this purpose, 'hybrid' measures which combine information from individual (e.g. FSM) and local neighbourhood (e.g. Acorn) proxies are preferable.

Of course, these findings should be interpreted considering the limitations of this research. First, our analysis has drawn comparisons between the proxies and permanent family income. However, the information available on family-income has been self-reported by cohort member's parents and may thus be subject to some measurement error. Likewise, we have also noted that the survey used suffers from a degree of non-response, which may reduce the representativeness of the sample. Second, the research has been conducted using data from England only, meaning the external validity of the results to other countries is not clear. Indeed, an important direction for future research is to understand how well the various proxies reflect family income differences across England, Northern Ireland, Scotland and Wales. For other countries, similar investigations to those presented in this paper are needed to understand the quality of family-income proxies available in administrative databases. Finally, we have found there to be merit in combining individual and local-neighbourhood information into a single, continuous proxy – particularly for researchers seeking to understand variation in outcomes at different points of the family income distribution. Yet future research, possibly using machine-learning techniques, is needed to better understand the optimum combination of variables to use when constructing such scales.

Despite these limitations, our findings have important implications for policy and practice. One is that government departments – such as the Department for Education in England – should construct for each pupil a score on a 'hybrid' scale that combines individual (e.g. FSM) and local neighbourhood (e.g. Acorn) information in the National Pupil Database. This, when used in conjunction with FSM eligibility, would provide analysts with a finer-grained proxy to explore differences in education outcomes across a broader array of family income groups.

Notes

1. In a working paper version, the authors construct the scale using a different set of measures (FSM, IMD and IDACI) noting that the first component explains 72% of the variation.
2. Polychoric PCA is used given that variables are included with different scales. An example of the Stata code used is.

```
Polychoricpca FSM IMD ACORN Neighbourhood_SES Neighbourhood_Ed
Neighbourhood_Tenure [pw = FOVWT1], score(IFS) nscore(1)
```
3. Where this information is not available at age 14, data from a previous wave is carried forward. For instance, say that IMD information is not available for a child at age 14; data is then used from the most recent previous survey wave (age 11) instead. As most measures we investigate are based on children's home postcode – which tends to be stable over time (unless their parents move house) – this is unlikely to have much impact on our results (other than slightly reducing variation in the proxies).
4. For instance, proxying poverty using the bottom IMD quintile may not be optimal, in the sense that it may not minimise the number of false-positives and maximise the number of true-positives. It may thus be better to cut the IMD distribution at a different point (e.g. the 30th percentile) instead.
5. As this package is unable to incorporate weights, we are unable to adjust for the complex MCS survey design in this part of our analysis.
6. This empirical approach requires the proxy to be continuous and ordinal. Acorn type has been treated as a continuous, ordinal measure for this purpose, though in reality this is likely to hold only approximately true. The OAC measure has been excluded due to this measure not clearly being even approximately ordinal.

7. Author's estimates using the MCS data. A similar issue occurs with the OAC measure which, using the University of Cambridge definition for university admissions, covers 38% of the population. See [Table 1](#) for further details.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

John Jerrim is a Professor of Education and Social Statistics at UCL Social Research Institute

References

- Acorn Technical Guide. 2017. UK Data Archive Study Number 8196 - Acorn Postcode-Level Directory for the United Kingdom. https://doc.ukdataservice.ac.uk/doc/8196/mrdoc/pdf/8196_acorn_technical_guide.pdf
- Adams, J., V. Ryan, and M. White. 2004. "How Accurate are Townsend Deprivation Scores as Predictors of Self-Reported Health? A Comparison with Individual Level Data." *Journal of Public Health* 27 (1): 101–106. <https://doi.org/10.1093/pubmed/fdh193>.
- Asaria, M., T. Doran, and R. Cookson. 2016. "The Costs of Inequality: Whole-Population Modelling Study of Lifetime Inpatient Hospital Costs in the English National Health Service by Level of Neighbourhood Deprivation." *Journal of Epidemiology and Community Health* 70 (10): 990–996. <https://doi.org/10.1136/jech-2016207447>.
- Atkinson, A., A. Maynard, and C. Trinder. 1983. "Evidence on Intergenerational Income Mobility in Britain. Some Further Preliminary Results." In *Human Resources, Employment and Development. International Economic Association Series*, edited by B. In: Weisbrod and H. Hughes (pp. 290–308). London: Palgrave Macmillan.
- Blanden, J., P. Gregg, and L. Macmillan. 2013. "Intergenerational Persistence in Income and Social Class: The Effect of Within-Group Inequality." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 176 (2): 541–563. <https://doi.org/10.1111/j.1467-985X.2012.01053.x>.
- Boliver, V., S. Gorard, and N. Siddiqui. 2021. "Using Contextual Data to Widen Access to Higher Education." *Perspectives: Policy and Practice in Higher Education* 25 (1): 7–13. <https://doi.org/10.1080/13603108.2019.1678076>.
- Breen, R., C. Mood, and J. Jonsson. 2015. "How Much Scope for a Mobility Paradox? The Relationship Between Social and Income Mobility in Sweden." *Sociological Science* 3:39–60. <https://doi.org/10.15195/v3.a3>.
- Broer, M., Y. Bai, and F. Fonseca. 2019. "A Review of the Literature on Socioeconomic Status and Educational Achievement." In *Socioeconomic Inequality and Educational Outcomes. IEA Research for Education (A Series of In-Depth Analyses Based on Data of the International Association for the Evaluation of Educational Achievement (IEA))*. (Vol. 5. pp 7–17). Hamburg, Germany: IEA.
- Bryere, J., C. Pernet, N. Copin, L. Launay, G. Gusto, P. Grosclaude, C. Delpierre, et al. 2017. "Assessment of the Ecological Bias of Seven Aggregate Social Deprivation Indices." *BMC Public Health* 17 (1): 86. <https://doi.org/10.1186/s12889-016-4007-8>.
- Bull, E., S. Dombrowski, N. McCleary, and M. Johnston. 2014. "Are Interventions for Low-Income Groups Effective in Changing Healthy Eating, Physical Activity and Smoking Behaviours? A Systematic Review and Meta-Analysis." *British Medical Journal Open* 4 (11): e006046. <https://doi.org/10.1136/bmjopen-2014-006046>.
- Burgess, S., C. Crawford, and L. Macmillan. 2018. "Access to Grammar Schools by Socio-Economic Status." *Environment & Planning A: Economy & Space* 50 (7): 1381–1385. <https://doi.org/10.1177/0308518X18787820>.

- Campbell, S., L. Macmillan, R. Murphy, and G. Wyness. 2019. "Inequalities in Student to Course Match: Evidence from Linked Administrative Data." Accessed March 31, 2020. <http://cep.lse.ac.uk/pubs/download/dp1647.pdf>.
- Charlton, J., C. Rudisill, N. Bhattarai, and M. Gulliford. 2013. "Impact of Deprivation on Occurrence, Outcomes and Health Care Costs of People with Multiple Morbidity." *Journal of Health Services Research & Policy* 18 (4): 215–223. <https://doi.org/10.1177/1355819613493772>.
- Chowdry, H., C. Crawford, L. Dearden, A. Goodman, and A. Vignoles. 2013. "Widening Participation in Higher Education: Analysis Using Linked Administrative Data." *Journal of the Royal Statistical Society: Series A* 176 (2): 431–457. <https://doi.org/10.1111/j.1467-985X.2012.01043.x>.
- Clayton, P. 2013. "CUTPT: Stata Module for Empirical Estimation of Cutpoint for a Diagnostic Test. Statistical Software Components S457719." Boston College Department of Economics. Accessed April 01, 2020. <https://ideas.repec.org/c/boc/bocode/s457719.html>.
- Connelly, R., V. Gayle, and P. S. Lambert (2016). A Review of Occupation-Based Social Classifications for Social Survey Research. *Methodological Innovations*, 9. <https://doi.org/10.1177/2059799116638003>
- Connelly, R., C. Playford, V. Gayle, and C. Dibben. 2016. "The Role of Administrative Data in the Big Data Revolution in Social Science Research." *Social Science Research* 59:1–12. <https://doi.org/10.1016/j.ssresearch.2016.04.015>.
- Crawford, C., L. Macmillan, and A. Vignoles. 2017. "When and Why Do Initially High-Achieving Poor Children Fall Behind?" *Oxford Review of Education* 43 (1): 88–108. <https://doi.org/10.1080/03054985.2016.1240672>.
- Currie, J. 2009. "Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development." *Journal of Economic Literature* 47 (1): 87–122. <https://doi.org/10.1257/jel.47.1.87>.
- Dika, S., and K. Singh. 2002. "Applications of Social Capital in Educational Literature. A Critical Synthesis." *Review of Educational Research* 72 (1): 31–60. <https://doi.org/10.3102/00346543072001031>.
- Education and Skills Funding Agency. 2018. "Schools Block Funding Formulae 2018-19: Analysis of Local authorities' Schools Block Funding Formulae." Accessed April 01, 2020. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/726783/Proforma_publication_18-19_FINAL_FOR_PUBLICATION.pdf.
- Galobardes, B., J. Lynch, and G. Davey-Smith. 2007. "Measuring Socioeconomic Position in Health Research." *British Medical Bulletin* 81-82 (1): 21–37. <https://doi.org/10.1093/bmb/ldm001>.
- Gorard, S., V. Boliver, N. Siddiqui, and P. Banerjee. 2019. "Which are the Most Suitable Contextual Indicators for Use in Widening Participation to HE?" *Research Papers in Education* 34 (1): 99–129. <https://doi.org/10.1080/02671522.2017.1402083>.
- Gorard, S., and N. Siddiqui. 2019. "How Trajectories of Disadvantage Help Explain School Attainment." *Sage Open* 9 (1): 215824401882517. <https://doi.org/10.1177/2158244018825171>.
- Gregg, P., L. Macmillan, and C. Vittori. 2017. "Moving Towards Estimating sons' Lifetime Intergenerational Economic Mobility in the UK." *Oxford Bulletin of Economics and Statistics* 79 (1): 79–100. <https://doi.org/10.1111/obes.12146>.
- Hajian-Tilaki, K. 2013. "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation." *Caspian Journal of Internal Medicine* 4 (2): 627–635.
- Hawkes, D., and I. Plewis 2008. "Missing Income Data in the Millennium Cohort Study: Evidence from the First Two Sweeps." *CLS working paper Working Paper 2008/10*.
- Hire, A., D. Ashcroft, D. Springate, and D. Steinke. 2018. "ADHD in the United Kingdom: Regional and Socioeconomic Variations in Incidence Rates Amongst Children and Adolescents (2004-2013)." *Journal of Attention Disorders* 22 (2): 134–142. <https://doi.org/10.1177/1087054715613441>.
- Hobbs, G., and A. Vignoles. 2010. "Is Children's Free School Meal 'Eligibility' a Good Proxy for Family Income?" *British Educational Research Journal* 36 (4): 673–690. <https://doi.org/10.1080/01411920903083111>.

- Jäntti, M., and S. Jenkins. 2015. "Income mobility." In *Handbook of Income Distribution*, edited by A. B. In: Atkinson and F. Bourguignon, 807–935. North Holland, Amsterdam, Holland: North-Holland.
- Lago, S., D. Cantarero, B. Rivera, M. Pascual, C. Blázquez-Fernández, B. Casal, and F. Reyes. 2018. "Socioeconomic Status, Health Inequalities and Non-Communicable Diseases: A Systematic Review." *Zeitschrift für Gesundheitswissenschaften Journal of public health* 26 (1): 1–14. <https://doi.org/10.1007/s10389-017-0850-z>.
- La Valle, I., L. Payne, E. Lloyd, and S. Potter. 2014. *Review of Policies and Interventions for Low-Income Families with Young Children—Summary Report*. London: Office of the Children's Commissioner
- Leibowitz, A. 1977. "Parental Inputs and Children's Achievement." *Journal of Human Resources* 12 (2): 242–251. <https://doi.org/10.2307/145387>.
- Link-Gelles, R., D. Westreich, A. E. Aiello, N. Shang, D. J. Weber, C. Holtzman, K. Scherzinger, et al. 2016. "Bias with Respect to Socioeconomic Status: A Closer Look at Zip Code Matching in a Pneumococcal Vaccine Effectiveness Study." *SSM - Population Health* 2:587–594. <https://doi.org/10.1016/j.ssmph.2016.08.005>.
- Ministry of Housing. 2019. "The English Indices of Deprivation 2019." Accessed September 20, 2022. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/835115/IoD2019_Statistical_Release.pdf.
- Mohammed, M., and A. Stevens. 2007. "The Value of Administrative Databases." *BMJ (Clinical Research Education)* 334 (7602): 1014–1015. <https://doi.org/10.1136/bmj.39211.453275.80>.
- Moore, J., L. Stinson, and E. Welniak. 2000. "Income Measurement Error in Surveys: A Review." *Journal of Official Statistics* 16 (4): 331–361.
- Office for Students. 2019. "Frequently Asked Questions About Area-Based Measures (POLAR and TUNDRA)." Accessed September 21, 2022. <https://www.officeforstudents.org.uk/media/3f1479d3-d144-4adb-b3c7-a6df6f996b27/polar-and-tundra-faqs-september-2019.pdf>.
- Office for Students. 2021. About the TUNDRA Area-Based Measures Data. Accessed September 20, 2022. <https://www.officeforstudents.org.uk/data-and-analysis/young-participation-by-area/about-tundra/>.
- Pampel, F., P. Krueger, and J. Denney. 2010. "Socioeconomic Disparities in Health Behaviors." *Annual Review of Sociology* 36 (1): 349–370. <https://doi.org/10.1146/annurev.soc.012809.102529>.
- Pattaro, S., N. Bailey, and C. Dibben. 2020. "Using Linked Longitudinal Administrative Data to Identify Social Disadvantage." *Social Indicators Research* 147 (3): 865–895. <https://doi.org/10.1007/s11205-019-02173-1>.
- Samson, L., K. Finegold, A. Ahmed, M. Jensen, C. Filice, and K. Joynt. 2017. "Examining Measures of Income and Poverty in Medicare Administrative Data." *Medical Care* 55 (12): e158–e163. <https://doi.org/10.1097/MLR.0000000000000606>.
- Sheringham, S., S. Sowden, M. Stafford, I. Simms, and R. Raine. 2009. "Monitoring Inequalities in the National Chlamydia Screening Programme in England: Added Value of ACORN, a Commercial Geodemographic Classification Tool." *Sexual Health* 6 (1): 57–62. <https://doi.org/10.1071/SH08036>.
- Social Mobility Commission. 2019. *State of the Nation 2018-19: Social Mobility in Great Britain*. London: England. Accessed April 01, 2020. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/798404/SMC_State_of_the_Nation_Report_2018-19.pdf.
- Soobader, M., F. LeClere, W. Hadden, and B. Maury. 2001. "Using Aggregate Geographic Data to Proxy Individual Socioeconomic Status: Does Size Matter?" *American Journal of Public Health* 91:632–636. <https://doi.org/10.2105/ajph.91.4.632>.
- Taylor, C. 2018. "The Reliability of Free School Meal Eligibility as a Measure of Socio-Economic Disadvantage: Evidence from the Millennium Cohort Study in Wales." *British Journal of Educational Studies* 66 (1): 29–51. <https://doi.org/10.1080/00071005.2017.1330464>.
- UNICEF Office of Research. 2018. *An Unfair Start: Inequality in Children's Education in Rich Countries*. Innocenti Report Card 15. UNICEF Office of Research – Innocenti: Florence.

- University of Cambridge. 2019. Access Agreement with the Office for Fair Access (OFFA) 2018-19. Accessed September 20, 2022. https://webcache.googleusercontent.com/search?q=cache:7ucqwZpAKk4J:https://www.undergraduate.study.cam.ac.uk/files/publications/university_of_cambridge_access_agreement_2018_19.pdf+%&cd=2&hl=en&ct=clnk&gl=uk.
- University of Oxford. 2022. "Disadvantage". Accessed September 20, 2022. <https://www.ox.ac.uk/about/facts-and-figures/admissions-statistics/undergraduate-students/current/disadvantage?wssl=1>.
- Wheeler, B. 2019. *Carstairs Index 2011 for Lower-Layer Super Output Areas*. 10.5255/UKDA-SN-851497. Colchester, Essex: UK Data Archive.
- Yousaf, S., and A. Bonsall. 2017. "UK Townsend Deprivation Scores from the 2011 Census Data." Accessed March 31, 2020. http://s3-eu-west-1.amazonaws.com/statistics.digitalresources.jisc.ac.uk/dkan/files/Townsend_Deprivation_Scores/UK%20Townsend%20Deprivation%20Scores%20from%202011%20census%20data.pdf.