Social Research Institute

UCL

Measuring socio-economic background using administrative data. What is the best proxy available?

John Jerrim

Quantitative Social Science
Working Paper No. 20-09
August 2020

Disclaimer

Quantitative Social Science
UCL Social Research Institute
University College London
55-59 Gordon Square
London WC1H 0NU

Measuring socio-economic background using administrative data. What is the best proxy available?

John Jerrim[1]

Abstract

Administrative data sources are increasingly being used to study socio-economic inequalities in education and health. Yet a well-known difficulty with such resources is the limited quality of information they hold about individual socio-economic position. Researchers, policymakers and practitioners using administrative data typically rely upon proxy indicators of individual socio-economic status, such as the Index of Multiple Deprivation (IMD) or eligibility for free school meals. But how well do these proxies actually capture socio-economic background? Relatively little existing work has considered this issue, with a particular dearth of studies drawing comparisons across the wide array of socio-economic status proxies now available. This study adds this evidence to the existing literature. Using a large, nationally-representative cohort study linked to administrative data, it is shown how eligibility for free school meals (averaged over the time a child has spent at school) is the best available proxy for childhood poverty, but is of limited use to researchers wanting to understand how key outcomes differ between young people from low, average and high socio-economic backgrounds. On the other hand, by combining individual and area level socio-economic proxies into a single continuous index, it is shown how administrative data can be used to produce robust estimates of family-income differences in key educational outcomes.

Keywords: Administrative data, proxy measures, socio-economic gaps, permanent income.
JEL Codes: I2

[1] Contact details: John Jerrim (J.Jerrim@ucl.ac.uk), UCL Social Research Institute,, University College London, 20 Bedford Way London, WC1H 0AL

1. Introduction

There has long been interest in the social and medical sciences in socio-economic inequalities in education and health (Lago et al 2018; Broer, Bai and Fonseca 2019). Numerous studies have documented how large socio-economic disparities across several dimensions emerge early in life (UNICEF 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 academic interest 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), with the Conservative party regularly stating how improving life chances for disadvantaged children is one of its key policy goals (Greening 2020).

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 (Pattaro, Bailey and Dibben 2020; Connelly et al 2016). A wide body of research in education and health in England has used resources such as the National Pupil Database (NPD), Health Episode Statistics (HES) and the Clinical Practise Research Datalink (CPRD) to add important new evidence on when and how socio-economic differences emerge, how they vary across different parts of the country and how they change over the life-course (Asaria, Doran and Cookson 2016; Hire et al 2018; Gorard and Siddiqui 2019). At the same time, such resources are also increasingly being used in policy and practise. For instance, key socio-economic variables available within administrative databases are now used to allocate funding to schools in England (Education and Skills Funding Agency 2018) and in making contextual offers to students applying to university (Gorard et al 2019).

However, one of the widely known 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, a data source would contain at least one of the three most commonly used individual indicators of socio-economic position (Galobardes, Lynch, Davey-Smith 2007):

- Social class measured by most recent occupation held
- Educational attainment
- (Permanent) family-income

Yet these key pieces of information are not routinely available in most administrative databases, meaning a proxy for socio-economic status must be used instead. Popular examples in England include eligibility for Free School Meals (FSM) and a raft of area-based measures such as the Index of Multiple Deprivation (IMD). Alternatively, sometimes individual and area-level socio-economic proxies are combined to produce a socio-economic status scale (Chowdry et al 2013) with the intuition being that this will provide more detailed information (and the best possible proxy) for individual socio-economic position. Yet relatively little is known about this collection of proxy measures, including how well they correlate with the actual underlying socio-economic measure of interest, whether they are subject to certain biases with respect to key demographic groups and the extent that their use in research can accurately capture the magnitude of socio-economic gaps (i.e. outcomes for the most advantaged and most disadvantaged socio-economic groups). Moreover, little work has considered how well a wide array of such possible measures perform against one another. This paper makes this contribution to the existing literature, investigating how well a wide-array of possible indicators proxy (a) permanent family income and (b) a multi-dimensional measure of socio-economic background (which combines information on parental education, social class and permanent family income).

There has of course been previous valuable work investigating the properties of socio-economic proxy measures, though these have typically considered a single indicator in isolation. Vignoles and Hobbs (2010) considered whether eligible for free school meals (FSM) in England is a good proxy for family income. They found that FSM children were more likely to be in low-income households, though this measure was unable to identify all low-income children within their dataset. Taylor (2018) conducted a similar investigation for FSM as a measure of low-income in Wales. He found that being eligible for FSM was a good proxy for socio-economic disadvantage, but that there is also a '*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 eligible for free school meals]'. Ilie, Sulivan and Vignoles (2017) investigate the properties of FSM and a set of area-based proxies (e.g. IMD, IDACI). They argue that neighbourhood-based measures are not as good as predicting educational achievement as FSM eligibility, and hence are not a good substitute. On the other hand, Crawford and Greaves (2013) find that FSM eligibility measured over three years successfully classifies about 80 percent of educationally disadvantaged young people, with other area-based socio-economic indicators (such as IDACI) performing almost as well. Investigating the most suitable contextual data to be used in university admissions, Gorard et al (2019) argue that FSM (and, in particular, the number of years a child has been eligible for FSM during their time at school) is one of the most suitable measures. 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 (FSM, Acorn, IMD), Campbell et al (2019) investigate in an appendix how well this measure correlates with the widely-used measure of occupational social class - finding a moderately strong correlation between the two. 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). It is argued that this neighbourhood index is strongly correlated with individual-level deprivation, and similarly predictive of health (at least when the index is calculated at the Census enumeration district level). 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 (in the context of the United States). The 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-Gellesa et al (2016) who found that "*zip code appears to be an adequate, though not perfect, proxy for individual SES*".

This paper seeks to contribute to this important literature in multiple ways. First, the properties of a wide array of socio-economic proxies are considered simultaneously, including individual-level indicators (e.g. FSM-eligibility), neighbourhood-level indicators (e.g. IMD) and 'hybrid' measures that combine elements of the two. 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 (or as a proxy for a multidimensional measure of socio-economic status), helping users of such indicators better understand their strengths and limitations. Third, most existing work has investigated how various socio-economic proxies correlate with a single measure of income gathered at one particular point in time. Yet this ignores a wide-ranging economic literature highlighting how it is *permanent* income that is likely to matter (Jantti and Jenkins 2015), with measurement error in the income data from a single year likely to be severe (Blanden, Gregg and Macmillan 2013). Finally, a unique contribution we make is to consider differences between 'transitory' (i.e. single point-in-time) and 'permanent' (i.e. time-averaged) socio-economic proxies, and how well they measure the underlying construct of interest (permanent family income or multidimensional SES). In other words, if we have a proxy measured across several years, does this enable the creation of a better indicator of individual socio-economic position?

To trail our key findings, we find eligibility for free school meals (averaged over the number of years children have spent at school) as the best available proxy for childhood poverty. On the other hand, two measures that are widely used as part of contextual admissions to England's universities (Polar and Tundra) stand out as particularly poor indicators of socio-economic background. None of the proxy measures we consider turn out to be a 'pure' indicator of socio-economic status, in that they all to some extent capture other characteristics of individuals (such as their ethnicity, housing tenure and family structure). Finally, for researchers trying to understand how key outcomes vary across the spectrum of the socio-economic status distribution, we show how there is much promise in a 'hybrid' measures that combines individual and area-level socio-economic proxies into a single continuous scale.

The paper now proceeds as follows. Section 2 describes the Millennium Cohort Study (MCS) dataset, with our empirical methodology following in section 3. Results are then presented in section 4, with discussion and conclusions following in section 5.

2. Data

The Millennium Cohort Study (MCS) is a rich, nationally-representative longitudinal study of UK children. A stratified, clustered survey design was used, with geographic areas (electoral wards) selected as the primary sampling unit, and then households with newly born children randomly selected from within (see Plewis 2004 for further details). Six sweeps have been conducted between 2000/01 and 2015, when children were nine months, 3, 5, 7, 11 and 14 years old. Parents, children and their teachers have been interviewed within the various sweeps. In total, 19,243 cohort members participated in the first survey, when children were nine months old (12,224 in England). Within this paper, we focus upon the MCS sample for England only, for cohort members whose data has been successfully linked to the National Pupil Database (NPD). This is due to many of the socio-economic proxies under investigation being country-specific (e.g. several area-based measures such as the Index of Multiple Deprivation are not designed to be comparable across England, Wales, Northern Ireland and Scotland). This leaves a final analytic sample of 7,439 individuals.

Access to a secure version of the MCS data was made available within a safe-setting by the Centre for Longitudinal Studies, as agreed with their Data Access Committee. This included information on cohort members' full postcodes across the first six MCS sweeps, along with selected data linked to the MCS from the NPD. This includes their eligibility for Free School Meals (FSM) for each year the child was registered at school and their GCSE grades. Such detailed information was necessary so that the area-based socio-economic proxies could be derived at the most fine-grained geographic level possible. Likewise, the linked NPD data allows us to draw comparisons between these area-based measures and FSM eligibility, which is currently the most widely-used socio-economic proxy used in education research, policy and practise in England.

Drawing upon this information, it is possible to derive a wide-array of socio-economic proxies that have previously been used in education, social and health research utilising administrative databases. A summary of these measures is provided in Table 1, with Appendix A providing further details about each, including information about how they have been derived.

<< Table 1 >>

Across the various MCS sweeps, children's parents have also been asked several questions about their income. Importantly, this not only included income from work but also from various other sources (e.g. benefits, investments) with information recorded for both mothers and fathers (where applicable). Hence, although the income data are self-reported – and unlikely to be entirely free from error (Moore, Stinson and Welniak 2000) – best practise has been followed in the collection of such information. Although there are some changes between waves, the survey organisers have used the information provided by respondents to harmonise the data across sweeps as far as possible. To derive the permanent income variable that is the focus of this paper, we take an average of total household income reported when the child was age nine months, 3, 5, 7, 11 and 14 (i.e. six separate time-points covering a 13-year period). Previous research has suggested that this is a sufficiently long horizon to provide a good measure of a family's permanent income (Gregg, Macmillan and Vittori 2017).

3. Methodology

3.1 Why focus upon permanent family income?

There are several different ways socio-economic status can be measured (Darin-Mattson, Fors and Kareholt 2017). This paper focuses upon proxies for (permanent) family income, relative poverty (permanently low income) and income-affluence (permanently high-income) for several reasons. First, income and poverty have long been of interest to both economists (Atkinson, Maynard and Trinder 1983) and sociologists (Breen, Mood and Jonsson 2016). Second, family income plays a central role in theoretical models of intergenerational persistence and how socio-economic inequalities in education, health and labour market outcomes are reproduced (Leibowitz 1977; Jerrim and Macmillan 2015). Third, much of the research and practical use of the proxies considered in this paper are (either implicitly or explicitly) as for an indicator of lack of familial resources (Crawford and Greaves 2013). Fourth, the concepts of income and poverty are widely understood – and discussed – amongst policymakers and practitioners, who often wish to target interventions upon low-income groups (La Valle et al 2014; Bull et al 2014). Finally, the alternatives to examining the relationship of each proxy with family income appear less attractive. For instance, there are several occupational measures of social class available (e.g. NS-SEC, International Socio-economic Index of Occupational Status), with it then debatable which of these specific measures should be used. Similarly, the other obvious alternative (parental education) suffers from the problem of the distribution changing significantly over time (Tasseva 2019).

For these reasons – and in order to give the paper a clear focus – our empirical analysis concentrates on how well each measure acts as a proxy for permanent family income. Nevertheless in Appendix B we also provide alternative results, 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). The pattern of results is, on the whole, qualitatively similar to those presented in the main text for permanent family income. We discuss the few exceptions where this is not the case.

3.2 Definitions

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. It is also possible to operationalise the notion of 'poverty' in different ways. Throughout this paper, we focus upon relative rather than absolute poverty, defined here as families in the bottom quintile of the MCS permanent income distribution.

Finally, although individuals with low-incomes are often the focus, research and policy are often also interested in outcomes for other income / socio-economic groups. For instance, academic research often tries to estimate the magnitude of socio-economic gaps (e.g. differences in academic achievement between children from the most and least advantaged backgrounds), while some key policy questions require information about individuals across the socio-economic continuum, including middle-income earners (Burgess, Crawford and Macmillan 2018). It is possible that a proxy measure might capture one part of the income distribution quite well, but not another. For instance, FSM eligibility might be a reasonable proxy for poverty, but provide little information to discriminate between middle and high income-earners. The empirical analysis therefore considers how well each measure proxies:

- Income poverty (permanently low-income). Defined as the bottom MCS permanent income quintile.

- Income affluence (permanently high-income). Defined as the top 20 percent of the MCS family-income quintile.
- Permanent income as a continuous variable.

3.3 Estimating the correlation between the proxy and permanent family income

One criterion for selecting a good proxy measure 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 poverty, affluence and a continuous measure of permanent income. Where the two measures being compared are both continuous (e.g. when comparing permanent income to the IMD) Pearson correlation coefficients will be presented. In contrast, point biserial correlations are estimated when one variable is categorical and the other continuous (e.g. when comparing FSM to permanent income) while polychoric correlations are used when both measures are categorical (e.g. when comparing FSM to income poverty).

Note that we estimate these correlations using both a 'transitory' measure of the proxy (taken as the value of the proxy measure when the cohort member was age 14[2]) and as a 'permanent' measure (an average across six time points spanning 13 years). This will reveal whether better proxy measures of permanent earnings can be derived if data is available (such as about home location) over a sustained period of time.

To further explore the association between the proxy measures and income poverty / affluence, 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 for each measure in Table 1 and Appendix A. We then calculate the 'true-positive' rate (e.g. the percent of cohort members the proxy measure correctly identifies as living in poverty) and the 'false-positive' rate (e.g. the percent of cohort members the proxy measure incorrectly identifies as living in poverty, when they do not) for each indicator in turn. Together, this approach helps illustrate the extent

[2] 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.

that each proxy can classify cohort members as coming from poor or affluent backgrounds correctly.

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 income affluence and the bottom fifth as a proxy for income poverty. These are then cross-tabulated with our 'true' measure of poverty as illustrated in Table 2. Figures in the top-left cell (shaded in green) provide the 'true positive' rate – the percent of children living in poverty that the proxy (bottom IMD quintile) correctly identifies as living in poverty. 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 living in poverty when they are not. With respect to the IMD, Table 2 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 various different proxies perform in this respect using a scatterplot, known formally as the Receiver Operating Characteristic (ROC) space (Hajian-Talaki 2013). This will be explained in further detail when presenting the results in the following section.

<< Table 2 >>

Of course, one limitation of this approach is that the percent of true-positive and false-positive cases for any given proxy depends upon 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 income poverty). 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)[3]. This is implemented via the Stata package cutpt (Clayton 2013). Appendix D provides information on the percentage of the population defined as 'advantaged' and 'disadvantaged' for each proxy when the optimal cut-point is used.

3.4 Investigating bias between key demographic groups (gender, ethnicity, single parent, renter/home owner, single parent, parental age group)

[3] 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 30^{th} percentile) instead.

The ideal proxy should also only capture the unobserved variable of interest (e.g. permanent family income) and not other characteristics of the individual, such as their gender, ethnicity, family structure or where they live. 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 individual socio-economic position). We explore potential bias in each measure as a proxy for permanent family income via estimation of 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 from this model 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 equal to 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 the β parameter, the greater the bias in the proxy measure. For instance, say that after controlling for the Index of Multiple Deprivation, we find there continues to be differences in permanent family income between young people of Black and White ethnic origin. This would indicate that the IMD partially captures the effect of ethnicity, rather than the underlying construct of interest (poverty/family income) alone.

The model outlined in equation (1) will be estimated separately for each proxy measure and each of the background measures listed below:

- 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)

Note that we are particularly interested in factors such as geographic region and homeownership given that several of the proxy measures are based upon local area-level characteristics (e.g. POLAR, IMD). Such area-based measures 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 will also replicate the analysis presented in equation (1) using a linear probability model, with our binary indicators of income-poverty and income-affluence as the outcome variables of interest. The proxy measure included as a control will also enter these models in its binary form (as documented in Table 1). These models will reveal whether certain groups are more (or less) likely to be categorised as 'poor' or 'affluent' by each proxy than would be the case if the true measure of interest (permanent family income in a categorised form) were available.

3.5 Estimating socio-economic achievement gaps

A final criterion for a good proxy measure is that it should have good predictive validity. We operationalise this concept within this analysis as the extent that each proxy can replicate permanent family income differences in academic achievement. In other words, the academic achievement of different socio-economic groups (as defined by each proxy) should be similar to those when using permanent family income. We therefore investigate the percent of 'disadvantaged' and 'advantaged' young people who achieved five A*-C grades in their GCSE examinations using each proxy measure, and how this compares to the analogous results when using permanent family income.

4. Results

Correlations

Table 3 begins by presenting the correlation between each proxy measure and (a) permanent income *poverty*; (b) permanent income and (c) permanent income *affluence*. Shading of cells should be read vertically, with green (red) cells indicating whether the correlation is stronger (weaker). The strength of the correlation is illustrated for the proxy measured at age 14 and when it is measured using an average value across all MCS sweeps (the ‘permanent’ rows). The correlation between age 14 family income and permanent family income is presented in the bottom row of Table 3 to help facilitate interpretation of results (i.e. it represents perhaps the best possible benchmark we could reasonably expect any of the proxy measures to meet).

<< Table 3 >>

There are four key points to note from Table 3. First, the Tundra and Polar measures that are widely used in England’s university admissions system are only weakly correlated with permanent income and (particularly) permanent income poverty; the correlation coefficients are around 0.2 to 0.4. Importantly, the results presented in Appendix B provide similar results when considering the correlation between Polar/Tundra and a multidimensional socio-economic status scale (which combines information on parental education, social class and income). This finding alone raises significant issues about interpreting Polar and Tundra as measures of socio-economic background, particularly in the context of contextualised university admissions. Indeed, it provides clear evidence that both are a worse measure of family background than many other viable alternatives.

Second, eligibility for FSM (particularly the ‘permanent’ measure that combines information from across several years) has the strongest correlation with permanent income poverty out of all the measures available. In fact, the correlation for time-average FSM and permanent income poverty (r = 0.69) is almost the same as between age 14 (‘transitory’) poverty and permanent poverty (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 proxy measures. This reflects the fact that FSM eligibility is a coarse indicator that only provides information about those towards the bottom end of the income distribution; it does not discriminate well between low, middle and high-income earners.

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

Finally, the only indicator for which there is a difference between the age 14 and time-average (permanent) versions of the proxy is for FSM. This likely reflects the fact that it is a binary indicator within a single year, with more information (and variation) to be gained when taking an average over time. Otherwise, in terms of the strength of the correlation with permanent income, it does not seem to matter if one takes an average value of the proxy over time, or just at one single point.

True-positive and false-positive rates

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

Where each proxy falls upon this graph is a function of two factors: (a) how well each proxy captures permanent income; (b) for continuous proxy measures, the threshold on the proxy measure below which a child will be classified as 'poor'. Two versions of the graph are therefore presented. Panel (a) on the left-hand side presents results where the grouping/cut-point used for each proxy to identify the 'in-poverty' group follows how the proxy measures are often used in research, policy and practise (for specific details about each indicator see Appendix A and Table 1). In contrast, panel (b) takes an empirical approach to determining the 'optimum' cut-point on the proxy to identify the disadvantaged group on the proxy measure (defined as the point along the proxy measure where the TP rate is maximised and the FP rate is minimised)[4]. Presenting both sets of estimates side-by-side helps illustrate why we obtain certain results.

<< Figure 1 >>

The first feature of note from Figure 1 is that it reaffirms the problems with interpreting Polar and Tundra as measures of socio-economic background. Both sit close to the dashed 45-degree line, with the true-positive rate almost identical to the false-positive rate. Clearly, these two

[4] 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 only approximately hold true. The OAC measure has been excluded due to this measure not clearly being even approximately ordinal.

measures are not good at identifying young people from low-income/disadvantaged socio-economic backgrounds (regardless of where the 'cut-point' is drawn to identify this group).

Acorn and the OAC proxies stand out in the left-hand panel (where conventional cut-points groupings to define the disadvantage group are used) 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 percent of children who genuinely live in poverty will be identified as socio-economically disadvantaged using the Acorn proxy (using the definition of 'disadvantage' applied by the University of Oxford in their contextual admissions criteria). Yet the false-positive rate – non-poor children that Acorn classifies as disadvantaged – also stands at around 40 percent (double that of most other measures considered). This, however, is a function of how the disadvantaged group using the Acorn measure has been defined; it actually encompasses half (49 percent) of the population[5]. As can be seen in panel (b), the true-positive and false-positive rate for Acorn is very similar to the other measures when the 'optimal' cut-point is used. This, nevertheless, highlights a key point when using proxy measures to identify (and target interventions at) particular socio-economic groups. It is not only important to consider how well the proxy captures the underlying construct of interest (such as permanently low family income) but also the 'cut-point' selected along the measure used to identify the most disadvantaged group.

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 on the measures are used (see panel a). This advantage disappears in panel (b), however, when the optimal cut-point is selected instead. Hence the advantage of the FSM measure in identifying children living in poverty 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 using each proxy measure to identify permanent income affluence can be found in Figure 2. Unsurprisingly, the average amount of time a child has been eligibile for

[5] 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 percent of the population. See Table Appendix A for further details.

FSM stands out from the other data points. It has a true-positive rate of almost one (due to almost all children from affluent backgrounds having never been eligible for FSM) but also a much higher false-positive rate (due to many children who have never been eligible for FSM not coming from particularly high-income backgrounds). Hence FSM is an outlier from the other indicators due to its coarse nature, with it not being well-suited to capturing differences between middle and high-income earners.

<< Figure 2 >>

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 (true-positive rate = 0.62) and the IFS measure (true-positive rate = 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 to determine income affluence, 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 (around 0.25) than most of the others (around 0.3). Together, this suggests that the Acorn and IFS measures may have some slight advantage at proxying income affluence 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.

Investigation of bias

Results considering the bias in each proxy as a measure of permanent income poverty can be found in Table 4. Analogous results for permanent income (as a continuous measure) and for permanent income affluence can be found in Appendix C. The figures refer to how much more likely the group in question (e.g. single versus two-parent households) are to actually live in permanent income poverty, having controlled for differences between groups in the proxy measure. Ideally, figures would be as close to zero as possible; this would indicate that the proxy captures all differences in income-poverty between the two groups. The shading of cells should be read vertically, with green cells indicating less bias in the proxy (i.e. values closer to zero) with red cells indicating where the bias is greater.

<< Table 4 >>

For many of the background characteristics considered, the absolute bias in each of the proxy measures is generally quite large. In other words, on this criterion, none of the proxy measures perform particularly well. Take the results for single parent households, for example. We find this group to generally be around 20 percentage points more likely to be living in income poverty than two-parent families, even after accounting for differences between these groups on the proxy measures. Similar results emerge in Appendix B, where we consider the bias in each proxy as an indicator of low (multi-dimensional) socio-economic status. This reveals how the proxies are not a 'pure' measure of permanent family income or of socio-economic background. They in part capture some other aspects about individual circumstances, such as ethnicity, housing tenure, family structure and (parental) age. The extent to which this matters is likely to depend upon the context in which the proxy is used.

Nevertheless, out of all the proxy indicators considered, permanent (time-averaged) FSM eligibility seems to outperform other measures on this criterion – 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 the FSM measure as an indicator of long-term poverty is much lower than for the other measures. Moreover, for none of the six characteristics examined does permanent FSM perform poorly relative to the other proxies available. Hence, in terms of selecting a single indicator of long-term poverty that has the least bias against key demographic groups, eligibility for FSM is likely to be the best pick.

Otherwise, there are two final features from Table 4 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 perform particularly poorly, not only with respect to proxying permanent income, but also low socio-economic status more generally (see Appendix B). In other words these two measures are also severely biased against certain demographic groups (if used or interpreted as a proxy for socio-economic background).

Estimation of achievement gaps

To conclude, Figure 3 investigates the magnitude of socio-economic achievement gaps using the various different proxy measures. The percent of teenagers from disadvantaged backgrounds (as defined by permanently low-income or by the proxy) who achieved five A*-C GCSE grades is plotted along the horizontal axis, with analogous figures for those from affluent backgrounds on the vertical axis. The square marker labelled ‘income’ presents the results for permanent income poverty/affluence (i.e. the ‘true’ income-achievement gap that the proxy measures should ideally replicate). As such, proxies that fall closer to the ‘income’ data point more closely replicate the desired results.

<< Figure 3 >>

It is immediately notable how the results for the FSM proxy 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 as a measure and its lack of flexibility. In particular, it does not discriminate well between middle and high-income earners, which limits its attractiveness as a measure to understand socio-economic differences in education and health outcomes – especially the advantages enjoyed by those from affluent backgrounds. This is an important (and often underappreciated) limitation of FSM as a proxy for socio-economic status.

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 groups it is underestimated. Take the results for the age 14 IMD, for example. Using this proxy, 37 percent of teenagers from disadvantaged backgrounds achieve five A*-C grades, compared to 70 percent of those from advantaged backgrounds (an income-achievement gap of 33 percentage points). This is compared to an 41 percentage 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.

Interestingly, this problem of attenuation seems to be greater in panel a of Figure 3 (using the age 14 measure of the proxy) compared to panel b (permanent, time-average of the proxy). In particular, most of the proxies move northwards in panel 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 socio-economic groups, there may be some benefit to researchers having access to proxies derived from address histories of individuals over a period of 10 – 15 years.

However, perhaps the standout feature of Figure 3 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 any of 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 socio-economic differences in key outcomes.

This point is reiterated by Figure 4, where we further probe this aspect of the IFS measure. 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 entire 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. Moreover, although the similarity of results when using the IFS proxy and the multidimensional measure of socio-economic status presented in Appendix B is slightly weaker, it still more accurately captures socio-economic achievement gaps than any of the alternatives.

<< Figure 4 >>

5. 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 socio-economic inequalities in education and health outcomes (Asaria, Doran and Cookson 2016), including socio-economic differences in access to health services (Charlton et al 2013) and family background differences in academic achievement (Gorard and Siddiqui 2019), including access to university (Chowdry et al 2013). One of the major limitations of such resources, however, is that they do not typically contain any of the most commonly used individual-level indicators of socio-economic background (social class, permanent income or education). Proxy measures of individual socio-economic status – many of which are measured at the local neighbourhood level – are used instead. Such proxy indicators are now frequently used in academic research (Ilie, Sutherland and Vignoles 2017) and by institutions making high-stakes decisions that will impact upon young people's lives (e.g. by universities when making offers to students).

Yet, despite their widespread use, relatively little is known about how well these proxy measures capture individual-level socio-economic position. There is a particular dearth of information in how they compare to one another in this respect. For instance, is there one proxy measure that consistently stands out as a better measure of socio-economic status than the others, and hence that academics and policymakers should always use? This paper has added this evidence to the literature, comparing the properties of a wide array of socio-economic proxy measures using data from a nationally representative, longitudinal database.

We find that the number of years a child has been eligible from FSM during their time at school is the best available marker for living in childhood poverty. Yet, as this proxy is focused upon the bottom end of the socio-economic status distribution, it is not well-suited to academic research (and public policy initiatives) attempting to understand the magnitude of socio-economic gaps (i.e. differences 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. Moreover, although no single indicator consistently stands out as 'better' than others, there are two (Polar and Tundra) that are clearly weaker measures of socio-economic position (and, particularly, socio-economic disadvantage).

Of course, these findings should be interpreted in light of the limitations of this research. First, our analysis has mainly drawn comparisons between the proxies and permanent family income (as well as a multidimensional measure of socio-economic status in Appendix B). 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. Second, the research has been conducted using data from England only, meaning the external validity of the results to other contexts (e.g. countries) is not clear. Indeed, an important direction for future research is to understand how well the various proxy measures capture socio-economic differences across England, Northern Ireland, Scotland and Wales. For other countries, similar investigations to those presented in this paper are needed, in order to understand the quality of socio-economic proxies available in administrative databases in other national settings. Third, although a wide-array of proxy variables have been considered, there remain some other proprietary measures (e.g. Mosaic) that it has not been possible to include due to access issues. Further investigation of such indicators is an important line of future enquiry. Finally, we have found there to be merit in combining individual and local-neighbourhood information into a single, continuous proxy scale – particularly for researchers seeking to understand variation in outcomes at different points of the socio-economic status distribution. Yet future research, possibly using machine-learning techniques, is needed to better understand the optimum combination of variables to use when constructing this scale.

Despite these limitations, our findings have a number of important implications for policy and practise. For instance, we recommend that the Department for Education in England stop releasing information on FSM eligibility to researchers accessing the National Pupil Database. Rather, they should construct for each pupil a 'hybrid' socio-economic scale score combining individual (e.g. FSM) and local neighbourhood (e.g. Acorn) information, and provide this to researchers instead. This would have the advantage of routinely providing academics with a finer-grained and more flexible indicator of children's socio-economic background, which we have shown to provide very similar estimates of achievement gaps to using permanent family-income. Moreover, as such 'hybrid' scales combine multiple pieces of information, it has much lower levels of sensitivity and disclosure risk than FSM eligibility or any of the individual area-based proxies.

Our other recommendation is that universities immediately stop using the Polar and Tundra measures as part of their contextual admissions process. This supports calls made by others who have previously voiced concerns over these indicators (Gorard et al 2019). In the short-term, universities can draw upon one of the many other neighbourhood proxies as an alternative, such as the IMD or Acord. (Although, we also advise that the regulator – the Office for Students – to provide clear guidance to universities as to how 'disadvantage' should be defined using these measures, that this is consistent across universities, and make clear the percentage of the population that fall into this 'disadvantaged' group). Longer-term, information on FSM-eligibility during a child's time at school would be the best information for universities to use in contextual admissions, given that it is already routinely collected by government and is the best available indicator for experience of long-term child poverty.

References

Adams, J.; Ryan, V. and White, M. 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.

Asaria, M.; Doran, T. and Cookson, R. 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 Epidemiol Community Health. doi:10.1136/jech-2016207447

Atkinson A.B., Maynard A.K., Trinder C.G. (1983) Evidence on Intergenerational Income Mobility in Britain: Some Further Preliminary Results. In: Weisbrod B., Hughes H. (eds) Human Resources, Employment and Development. International Economic Association Series. Palgrave Macmillan, London.

Blanden, J., Gregg, P., & Macmillan, L. (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.

Breen, R.; Mood, C. and Jonsson, J. 2016. How Much Scope for a Mobility Paradox? The Relationship between Social and Income Mobility in Sweden. *Sociological Science.* DOI 10.15195/v3.a3

Broer M., Bai Y., Fonseca F. (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. Springer, Cham

Bryere, J.; Pornet, C.; Copin, N. et al. 2017. Assessment of the ecological bias of seven aggregate social deprivation indices. *BMC Public Health* 17**:** 86. https://doi.org/10.1186/s12889-016-4007-8

Bull, E.; Dombrowski S.; McCleary, N. and Johnston, M. 2014 Are interventions for low-income groups effective in changing healthy eating, physical activity and smoking behaviours? A systematic review and meta-analysis *BMJ Open*: 2014;4:e006046. doi: 10.1136/bmjopen-2014-006046

Burgess, S.; Crawford, C.; and Macmillan, L. 2018. Access to grammar schools by socio-economic status. Accessed 31/03/2020 from https://discovery.ucl.ac.uk/id/eprint/10053242/1/Burgess,%20Crawford%20and%20Macmillan_GrammarandSES_Revised.pdf

Charlton, J., Rudisill, C., Bhattarai, N. and Gulliford, M. (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.; Crawford, C.; Dearden, L.; Goodman, A. and Vignoles, A. 2013. Widening participation in higher education: analysis using linked administrative data. *Journal of the Royal Statistical Society Series A*, 176: 431-457.

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 01/04/2020 from https://ideas.repec.org/c/boc/bocode/s457719.html

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

Crawford, C. and Greaves, E. 2013. A comparison of commonly used socio-economic indicators: their relationship to educational disadvantage and relevance to Teach First. IFS report R79. Accessed 01/04/2020 from https://www.ifs.org.uk/comms/r79.pdf

Crawford, C.; Macmillan, L. and Vignoles, A. (2017). When and why do initially high-achieving poor children fall behind? *Oxford Review of Education* 43(1): 88-108, DOI: 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. Retrieved April 1, 2020, from www.jstor.org/stable/27647135

Darin-Mattsson, A., Fors, S. & Kåreholt, I. Different indicators of socioeconomic status and their relative importance as determinants of health in old age. *Int J Equity Health* **16,** 173 (2017). https://doi.org/10.1186/s12939-017-0670-3

Education and Skills Funding Agency. 2018. Schools block funding formulae 2018-19: Analysis of local authorities' schools block funding formulae. Accessed 01/04/2020 from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/726783/Proforma_publication_18-19_FINAL_FOR_PUBLICATION.pdf

Gorard, S., & Siddiqui, N. (2019). How Trajectories of Disadvantage Help Explain School Attainment. *SAGE Open*. https://doi.org/10.1177/2158244018825171

Gorard, S.; Boliver, V.; Siddiqui, N. and Banerjee, P. 2019. Which are the most suitable contextual indicators for use in widening participation to HE? *Research Papers in Education* 34(1): 99-129.

Gregg, P., Macmillan, L. and Vittori, C. (2017), Moving Towards Estimating Sons' Lifetime Intergenerational Economic Mobility in the UK. Oxf Bull Econ Stat, 79: 79-100. doi:10.1111/obes.12146

Greening, J. (2020). Boris Johnson's social mobility plan could create the change we've been waiting for – now he just has to deliver. *The Independent*, Wednesday 8th January 2020. Accessed 01/04/2020 from https://www.independent.co.uk/voices/social-mobility-conservative-government-boris-johnson-justine-greening-a9275686.html

Galobardes, B.; Lynch, J. and Davey-Smith, G. 2007. Measuring socioeconomic position in health research *British Medical Bulletin* 81-82(1): 21–37.

Hajian-Tilaki K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian journal of internal medicine*, *4*(2): 627–635.

Hire, A. J., Ashcroft, D. M., Springate, D. A., & Steinke, D. T. (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 Vignoles, A. 2010. Is children's free school meal 'eligibility' a good proxy for family income? *British Educational Research Journal* 36(4): 673-690.

Ilie, S., Sutherland, A. and Vignoles, A. 2017. Revisiting free school meal eligibility as a proxy for pupil socio-economic deprivation. *British Educational Research Journal* 43(2): 253-274. doi:10.1002/berj.3260

Jäntti, Markus and Jenkins, Stephen P. (2015). *Income mobility.* In: Atkinson, Anthony B. and Bourguignon, François, (eds.) Handbook of Income Distribution. North Holland, Amsterdam, Holland, pp. 807-935.

Jerrim, J. and Macmillan, L. 2015. Income inequality, intergenerational mobility and the Great Gatsby Curve: is education the key? *Social Forces*, 94(2), 505-533.

La Valle, I; Payne, L.; Lloyd, E. and Potter, S.(2014). Review of policies and interventions for low-income families with young children–Summary report. London: Office of the Children's Commissioner

Lago, S., Cantarero, D., Rivera, B., Pascual, M., Blázquez-Fernández, C., Casal, B., & Reyes, F. (2018). Socioeconomic status, health inequalities and non-communicable diseases: a systematic review. *Zeitschrift fur Gesundheitswissenschaften = Journal of public health*, *26*(1), 1–14. https://doi.org/10.1007/s10389-017-0850-z

Leibowitz, Arleen. 1977. Parental Inputs and Children's Achievement. *Journal of Human Resources* 12(2):242–51.

Lewis-Beck, M. S., Bryman, A. and Futing Liao, T. (2004). *The SAGE encyclopedia of social science research methods* (Vols. 1-0). Thousand Oaks, CA: Sage Publications, Inc. doi: 10.4135/9781412950589

Link-Gelles, R., Westreich, D., Aiello, A. E., Shang, N., Weber, D. J., Holtzman, C., Scherzinger, K., Reingold, A., Schaffner, W., Harrison, L. H., Rosen, J. B., Petit, S., Farley, M., Thomas, A., Eason, J., Wigen, C., Barnes, M., Thomas, O., Zansky, S., Beall, B., … Moore, M. R. (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

Mohammed, M. and Stevens, A. (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.; Stinson, L. and Welniak, E. 2000. Income measurement error in surveys: A review. *Journal of Official Statistics* 16(4): 331-361.

Pampel, F. C., Krueger, P. M., & Denney, J. T. (2010). Socioeconomic Disparities in Health Behaviors. *Annual review of sociology*, *36*, 349–370. https://doi.org/10.1146/annurev.soc.012809.102529

Pattaro, S., Bailey, N. & Dibben, C. 2020. Using Linked Longitudinal Administrative Data to Identify Social Disadvantage. *Social Indicators Research* 147, 865–895. https://doi.org/10.1007/s11205-019-02173-1

Samson, L.; Finegold, K.; Ahmed, A.; Jensen, M.; Filice, C. and Joynt, K. 2017. Examining Measures of Income and Poverty in Medicare Administrative Data. Medical Care. 55(12): e158–e163.

Sheringham, S.; Sowden, S.; Stafford, M.; Simms, I. and Raine, R. 2009. Monitoring inequalities in the National Chlamydia Screening Programme in England: added value of ACORN, a commercial geodemographic classification tool. *Sexual Health* 6: 57-62.

Social Mobility Commission. 2019. State of the Nation 2018-19: Social Mobility in Great Britain. London: England. Accessed 01/04/2020 from 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.; LeClere, F.; Hadden, W. and Maury, B. 2001. Using Aggregate Geographic Data to Proxy Individual Socioeconomic Status: Does Size Matter? *American Journal of Public Health* 91: 632-636.

Tasseva, I. 2019. The Changing Education Distribution and Income Inequality in Great Britain. Euromod working paper series EM16/19. Accessed 01/04/2020 from https://www.euromod.ac.uk/sites/default/files/working-papers/em16-19.pdf

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, DOI: 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.

Table 1. The socio-economic proxy measures investigated in this paper

Measure	Level measured at	Permanent measure	Proxy for poverty		Proxy for affluence	
			Definition	% in group	Definition	% in group
Index of Multiple Deprivation	LSOA	Mean over time	Bottom quintile	20%	Top quintile	20%
ACORN	Postcode	Mode over time	Category 4 / 5	49%	Category 1 /2	23%
Free-school meals	Individual	-	-	17%	-	83%
Years of free school meals	Individual	Mean over time	% time FSM eligible	20%	Never FSM	67%
Income Deprivation Affecting Children	LSOA	Mean over time	Bottom quintile	20%	Top quintile	20%
Carstairs Index	LSOA	Mean over time	Bottom quintile	20%	Top quintile	20%
Output Area Classification	LSOA	Mode over time	Groups 7,8 3a-3c, 4b	38%	See Appendix A	18%
IFS socio-economic status index	Individual / postcode	Mean over time	Bottom quintile	20%	Top quintile	20%
Townsend index	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	MSOA	Mean over time	Bottom quintile	21%	Top quintile	20%
Transitory income (age 14)	Individual	N/A	Bottom quintile	20%	Top quintile	20%

Notes: OA = output area; LSOA = Lower super output area; MSOA = middle super output area.

Table 2. An illustration of how the true-positive and false-positive rate for income poverty is calculated using the Index of Multiple Deprivation (IMD).

	In poverty	Not in poverty
Most disadvantaged IMD quartile	43.7%	14.1%
Not in the most disadvantaged 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).

Table 3. The correlation between different proxy measures and permanent family income

Measure	Type	Correlation with poverty	Correlation with permanent income	Correlation with affluence
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

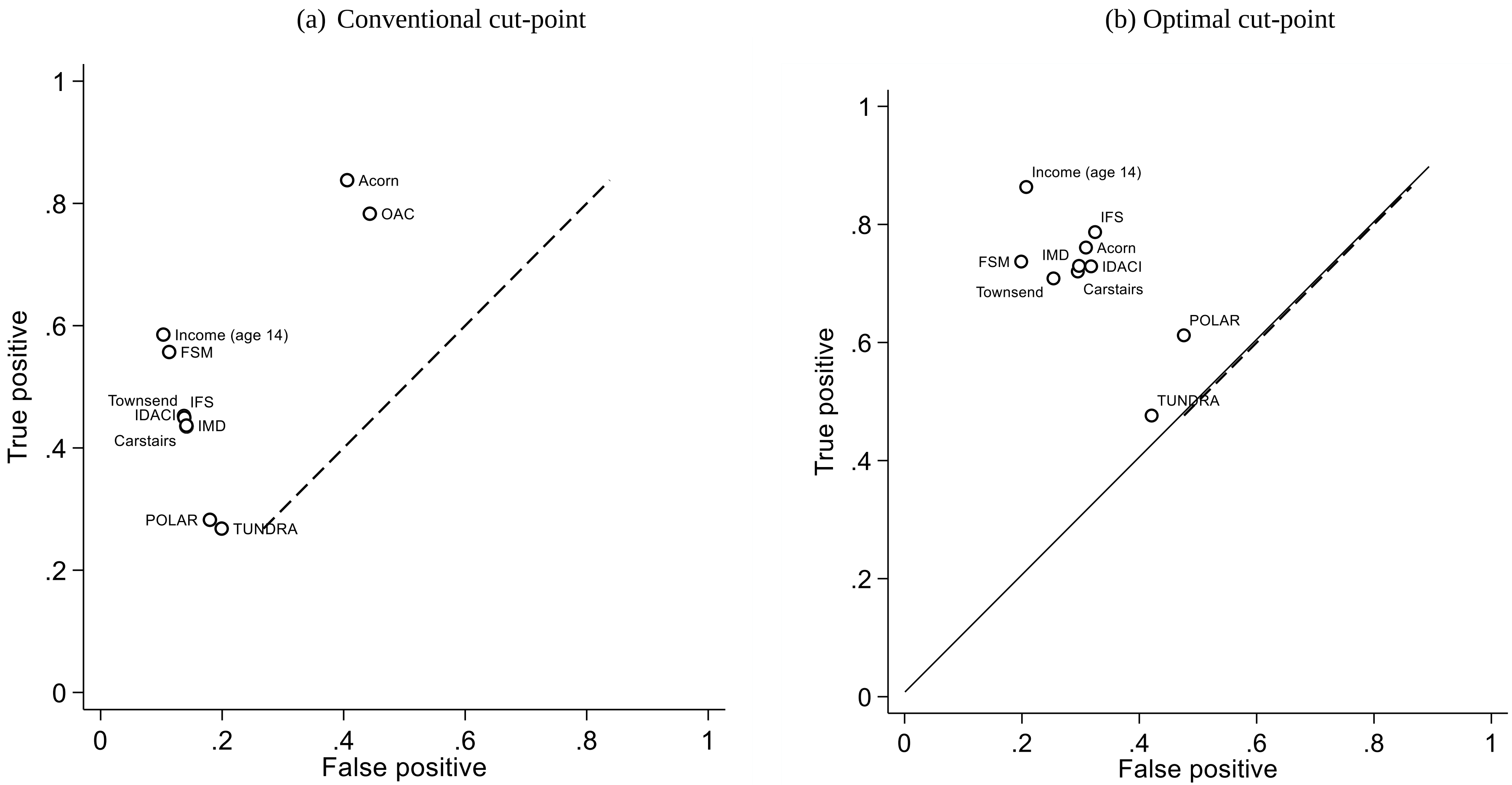
Notes: Shading should be read vertically. Higher correlations are in green shades; lower correlations in red shades.

Table 4. 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%

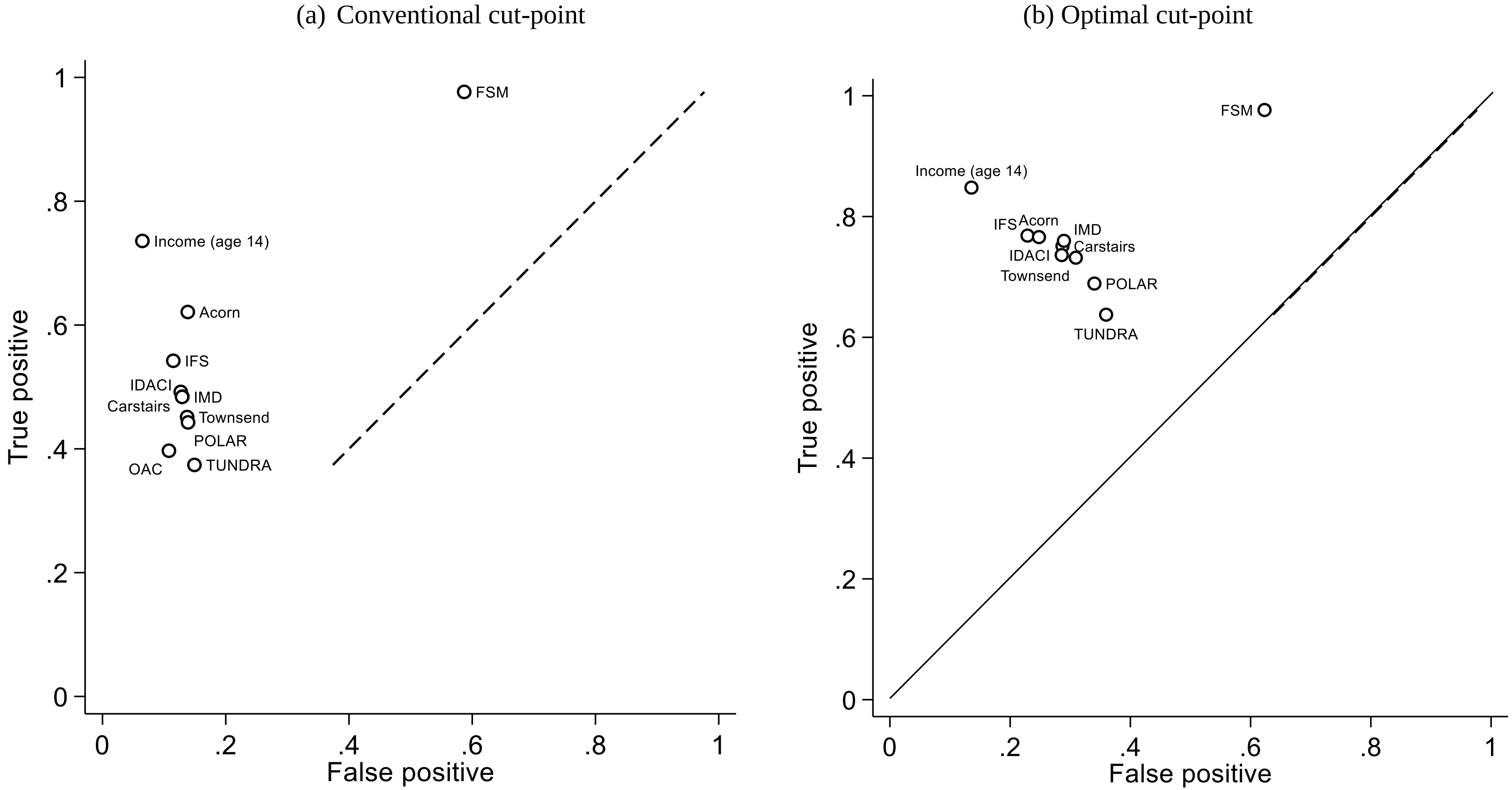
Notes: Figures indicate how much more likely the group is have permanently low-income, conditional upon the proxy measure. For instance, Londoners are around 14 percentage points more likely to actually live in income poverty than those living elsewhere in England, conditional upon the age 14 POLAR proxy 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).

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



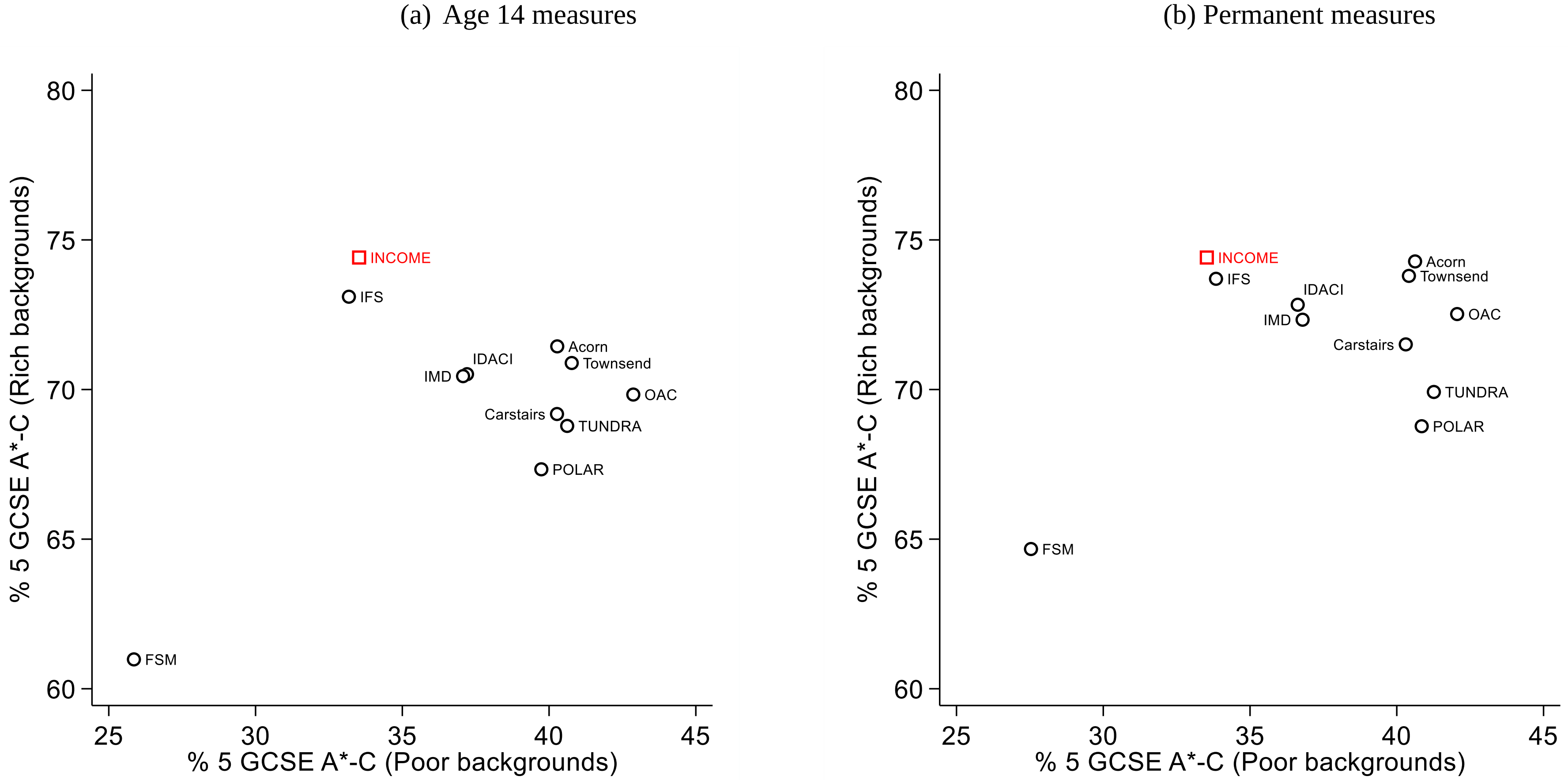
Notes: True-positive rate is presented on the vertical axis; this captures the percent of low-income families correctly identified by the proxy. The false-positive rate is plotted along the horizontal axis, capturing the percent of families the proxy identifies as coming from a low-income background 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.

Figure 3. True-positive and false-positive rates for detecting income affluence using different proxy measures



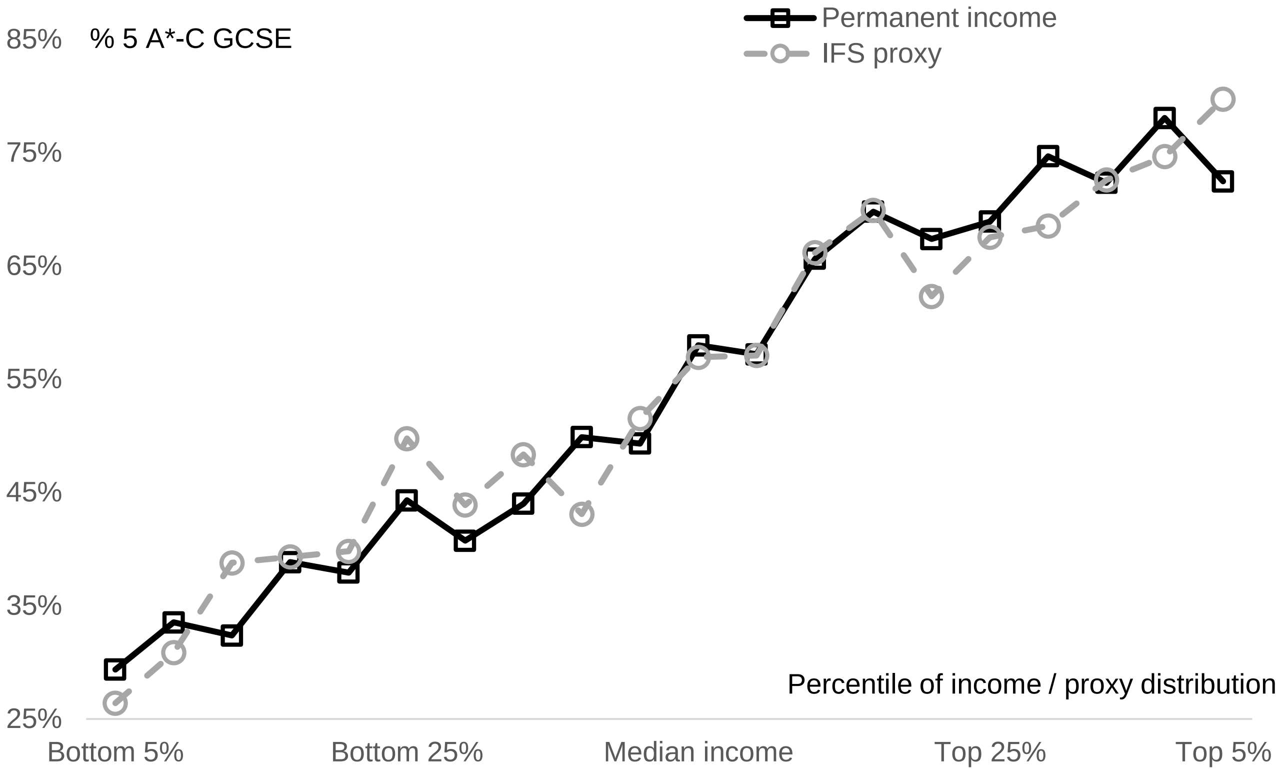
Notes: True-positive rate is presented on the vertical axis; this captures the percent of high-income families correctly identified by the proxy. The false-positive rate is plotted along the horizontal axis, capturing the percent of families the proxy identifies as coming from a high-income background 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.

Figure 4. How well do the proxy measures capture permanent family income gaps (top versus bottom quintile) in academic achievement? (Conventional cut-points)



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.

Figure 5. 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.

Appendix A. An overview of the socio-economic proxies investigated in this paper

Acorn

Overview

Acorn is a geodemographic classification system developed by CACI Limited. The data are proprietary, though can be accessed for academic research purposes from the UK Data Service (SN8196 - https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8196). The Acorn classification system combines information from the Land Registry, administrative data and commercial data to divide each postcode in the UK into one of six Acorn categories, 18 Acorn groups and 62 Acorn types, as shown in the table below (a 'pen-portrait' of each Acorn type is available from https://acorn.caci.co.uk/downloads/Acorn-User-guide.pdf). These 62 Acorn types are based upon a combination of information and data sources, such as house sales, house rentals, accommodation designed for elderly people, high rise social housing, other housing lists, care accommodation, student accommodation, information about residents, benefits claimants, census and lifestyle data.

Within this paper, we treat the 62 Acorn types as an ordinal proxy of socio-economic status. The Acorn data used in this paper is available at the postcode level; postcodes in the UK have a median size of 13 households and 31 residents (Acorn Technical Guide 2017). Further details about the methodology underpinning the Acorn methodology are available from http://doc.ukdataservice.ac.uk/doc/8196/mrdoc/pdf/8196_acorn_technical_guide.pdf. The full postcode of recorded for children in the MCS is used to assign them to Acorn categories, groups and types.

We follow the University of Oxford outreach team to create our measure of disadvantage using the Acorn classification system (see https://www.ox.ac.uk/about/facts-and-figures/admissions-statistics/undergraduate-students/current/disadvantage?wssl=1). Specifically, 'disadvantaged' is defined as any child in Acorn Category 4 – Financially stretched (excluding Type 34, student flats and halls of residence) – and Acorn Category 5 – Urban Diversity. Our analysis of the MCS suggests that approximately 49 percent of the population fall into one of these groups (and hence meet the Acorn definition of 'disadvantaged' used in this paper). When treating Acorn Type as an ordinal measure, we find the 'optimal' cut point to define disadvantage as including Type 40 and above (with approximately 41 percent of the population then falling into the disadvantaged group).

Table B1. The structure of the Acorn classification system

Category	Group	Type
1. Affluent achievers	A Lavish Lifestyles	1. Exclusive enclaves
		2. Metropolitan money
		3. Large house luxury
	B Executive Wealth	4. Asset rich families
		5. Wealthy countryside commuters
		6. Financially comfortable families
		7. Affluent professionals
		8. Prosperous suburban families
		9. Well-off edge of towners
	C Mature Money	10. Better-off villager
		11. Settled suburbia, older people
		12. Retired and empty nesters
		13. Upmarket downsizers

2. Rising prosperity	D City Sophisticates	14. Townhouse cosmopolitans 15. Younger professionals in smaller flats 16. Metropolitan professionals 17. Socialising young renters
	E Career Climbers	18. Career driven young families 19. First time buyers in small, modern homes 20. Mixed metropolitan areas
3. Comfortable communities	F Countryside Communities	21. Farms and cottages 22. Larger families in rural areas 23. Owner occupiers in small towns and villages
	G Successful Suburbs	24. Comfortably-off families in modern housing 25. Larger family homes, multi-ethnic areas 26. Semi-professional families, owner occupied neighbourhoods
	H Steady Neighbourhoods	27. Suburban semis, conventional attitudes 28. Owner occupied terraces, average income 29. Established suburbs, older families
	I Comfortable Seniors	30. Established suburbs, older families 31. Elderly singles in purpose-built accommodation
	J Starting Out	32. Educated families in terraces, young children 33. Smaller houses and starter homes
4. Financially stretched	K Student Life	34. Student flats and halls of residence 35. Term-time terraces 36. Educated young people in flats and tenements
	L Modest Means	37. Low cost flats in suburban areas 38. Semi-skilled workers in traditional neighbourhoods 39. Fading owner occupied terraces 40. High occupancy terraces, many Asian families
	M Striving Families	41. Labouring semi-rural estates 42. Struggling young families in post-war terrace 43. Families in right-to-buy estates 44. Post-war estates, limited means
	N Poorer Pensioners	45. Pensioners in social housing, semis and terrace 46. Elderly people in social rented flats 47. Low income older people in smaller semis 48. Pensioners and singles in social rented flats
5. Urban adversity	O Young Hardship	49. Young families in low cost private flats 50. Struggling younger people in mixed tenure 51. Young people in small, low cost terraces
	P Struggling Estates	52. Poorer families, many children, terraced housing 53. Low income terraces 54. Multi-ethnic, purpose-built estates 55. Deprived and ethnically diverse in flats 56. Low income large families in social rented semis
	Q Difficult Circumstances	57. Social rented flats, families and single parents 58. Singles and young families, some receiving benefits 59. Deprived areas and high-rise flats
6. Not private households	R Not Private Households	60. Active communal population 61. Inactive communal population 62. Business addresses without resident population

Notes: Green (red) shading indicates the groups included in the advantaged (disadvantaged) groups, using conventional cut-offs.

‘Advantaged’ is defined as any child in Acorn Category 1 (affluent achievers) and Category 2 (rising prosperity). Our analysis of the MCS suggests that approximately 23 percent of the population fall into one of these groups (and hence meet the Acorn definition of ‘advantaged’ used in this paper). When treating Acorn Type as an ordinal measure, we find the ‘optimal’ cut point to define disadvantage as including Type 1 to Type 26 (with approximately 35 percent of the population then falling into the advantaged group).

Output Area Classification (OAC)

Overview

The OAC is a geodemographic classification system developed by the Office of National Statistics. The data are open source, with each census Output Area (which is comprised of around 125 households) being classified into one of eight OAC groups, 26 groups and 76 sub-groups, as shown in the table below (a 'pen-portrait' describing these groupings is available from https://webcache.googleusercontent.com/search?q=cache:hQS3UR5PDFwJ:https://www.ons.gov.uk/file%3Furi%3D/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications/penportraitsandradialplots/penportraits.pdf+&cd=2&hl=en&ct=clnk&gl=uk). The OAC data are categorical – and not clearly ordinal. These groupings have been formed based upon the demographic structure (e.g. age, marital status, ethnicity), household composition, housing type (e.g. detached house, flats, property ownership), socio-economic (e.g. educational qualifications, car ownership) and employment situation (e.g. industry of occupation, percentage of people in employment) of the output area. Further details about how the groupings have been formed are available from https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications/methodologyandvariables

Within the MCS, the census output area in which children live is used to assign them to OAC supergroups, groups and subgroups. The OAC classification for each output area can be accessed from https://www.ons.gov.uk/methodology/geography/geographicalproducts/areaclassifications/2011areaclassifications/datasets

Measurement of disadvantage / advantage

We follow the University of Cambridge outreach team to create our measure of disadvantage using the OAC (see 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). In particular, as part of their contextual admissions, they define disadvantage as encompassing the following groups:

- Subgroup 3a1 – Established renting families
- Subgroup 3a2 – Young families and students
- Subgroup 3b1 – Striving service workers
- Subgroup 3b2 – Bangladeshi mixed employment
- Subgroup 3b3 – Multi-ethnic professional service workers
- Subgroup 3c1 – Constrained neighbourhoods
- Subgroup 3c2 – Constrained commuters
- Subgroup 4a1 – Social renting young
- Subgroup 4a2 – Private renting new
- Subgroup 4b1 – Asian terraces and flats
- Subgroup 4b2 – Pakistani communities
- Subgroup 4c2 – Multicultural new arrivals
- Subgroup 6b3 – Semi-detached aging
- Group 7a –Challenged diversity
- Group 7b – Constrained flat dwellers
- Group 7c – White communities
- Super-group 8 – Hard pressed living

Table B2. Structure of the OAC groupings

Super group	Groups	Subgroups
1. Rural residents	1a. Farming communities	1a1 Rural workers and families
		1a2 – Established farming communities
		1a3 – Agricultural communities
		1a3 – Agricultural communities
		1a4 – Older farming communities
	1b – Rural tenants	1b1 – Rural life
		1b2 – Rural white-collar workers
		1b3 – Ageing rural flat tenants
	1c – Ageing rural dwellers	1c1 – Rural employment and retiree
		1c2 – Renting rural retirement
		1c3 – Detached rural retirement
2. Cosmopolitans	2a – Students around campus	2a1 – Student communal living
		2a2 – Student digs
		2a3 – Students and professionals
	2b – Inner city students	2b1 – Students and commuters
		2b2 – Multicultural student neighbourhoods
	2c – Comfortable cosmopolitan	2c1 – Migrant families
		2c2 – Migrant commuters
		2c3 – Professional service cosmopolitans
	2d – Aspiring and affluent	2d1 – Urban cultural mix
		2d2 – Highly-qualified quaternary workers
		2d3 – EU white-collar workers
3. Ethnicity central	3a – Ethnic family life	3a1 – Established renting families
		3a2 – Young families and students
	3b - Endeavouring Ethnic Mix	3b1 – Striving service workers
		3b2 – Bangladeshi mixed employment
		3b3 – Multi-ethnic professional service workers
	3c – Ethnic dynamics	3c1 – Constrained neighbourhoods
		3c2 – Constrained commuters
	3d – Aspirational techies	3d1 – New EU tech workers
		3d2 – Established tech workers
		3d3 – Old EU tech workers
4. Multicultural metropolitans	4a – Rented family living	4a1 – Social renting young families
		4a2 – Private renting new arrivals
		4a3 – Commuters with young families
	4b – Challenged Asian terraces	4b1 – Asian terraces and flats
		4b2 – Pakistani communities
	4c – Asian traits	4c1 – Achieving minorities
		4c2 – Multicultural new arrivals
		4c3 – Inner city ethnic mix
5. Urbanites	5a – Urban professionals and families	5a1 – White professionals
		5a2 – Multi-ethnic professionals with families
		5a3 – Families in terraces and flats
	5b – Ageing urban living	5b1 – Delayed retirement
		5b2 – Communal retirement
		5b3 – Self-sufficient retirement
6. Suburbanites	6a – Suburban achievers	6a1 – Indian tech achievers
		6a2 – Comfortable suburbia
		6a3 – Detached retirement living
		6a4– Ageing in suburbia

	6b – Semi-detached suburbia	6b1 – Multi-ethnic suburbia 6b2 – White suburban communities 6b3 – Semi-detached ageing 6b4 – Older workers and retirement
7. Constrained city dwellers	7a – Challenged diversity	7a1 – Transitional Eastern European neighbourhoods 7a2 – Hampered aspiration 7a3 – Multi-ethnic hardship
	7b – Constrained flat dwellers	7b1 – Eastern European communities 7b2 – Deprived neighbourhoods 7b3 – Endeavouring flat dwellers
	7c – White communities	7c1 – Challenged transitionaries 7c2 – Constrained young families 7c3 – Outer city hardship
	7d – Ageing city dwellers	7d1 – Ageing communities and families 7d2 – Retired independent city dwellers 7d3 – Retired communal city dwellers 7d4 – Retired city hardship
8. Hard-pressed living	8a – Industrious communities	8a1 – Industrious transitions 8a2 – Industrious hardship
	8b – Challenged terraced workers	8b1 – Deprived blue-collar terraces 8b2 – Hard pressed rented terraces
	8c –Hard pressed ageing workers	8c1 – Ageing industrious workers 8c2 – Ageing rural industry workers 8c3 – Renting hard-pressed worker
	8d – Migration and churn	8d1 – Young hard-pressed families 8d2 – Hard-pressed ethnic mix 8d3 – Hard-Pressed European Settlers

Notes: Cells shaded in red indicate the definition of the OAC 'disadvantaged' group.

Our analysis of the MCS suggests that approximately 18 percent of the population fall into one of these groups (and hence meet the OAC definition of 'disadvantaged' used in this paper). As the OAC classification system is not clearly ordinal, it is not possible to calculate the 'optimum' cut-point calculated for the other socio-economic proxies investigated in this paper.

As far as we are aware, there is not commonly used measure of 'advantage' used for the OAC. In the paper, we have therefore simply used the MCS to establish the OAC sub-groups with the highest average weekly pay. Our advantaged classification then takes those sub-groups that together account for the top 20 percent of the permanent family income distribution. The OAC definition of advantage (which captures approximately the mot advantaged 20 percent of the population) is hence based upon the following OAC sub-groups:

- 1a2 – Established farming communities
- 1b2 – Rural white-collar workers
- 1c1 – Rural employment and retiree
- 1c3 – Detached rural retirement
- 2a3 – Students and professionals
- 2c2 – Migrant commuters

- 2c3 – Professional service cosmopolitans
- 2d1 – Urban cultural mix
- 2d2 – Highly-qualified quaternary workers
- 2d3 – EU white-collar workers
- 3d2 – Established tech workers
- 4c3 – Inner city ethnic mix
- 5b1 – Delayed retirement
- 5b2 – Communal retirement
- 6a1 – Indian tech achievers
- 6a2 – Comfortable suburbia
- 6a3 – Detached retirement living
- 6a4– Ageing in suburbia

Index of Multiple Deprivation (IMD)

The Index of Multiple Deprivation (IMD) is the official measure of relative deprivation used in England. It is comprised of seven deprivation domains, measured at Lower Super Output Area (LSOA) level, that are combined (with unequal weight) to form the final scale. This includes:

- Income (22.5% weight)
- Employment (22.5% weight)
- Health deprivation and disability (13.5% weight)
- Education, skills and training (13.5% weight)
- Crime (9.3% weight)
- Barriers to Housing and services (9.3% weight)
- Living environment (9.3% weight)

In total, the IMD combines information from across 39 separate indicators, most of which are sourced from government administrative data sources (although information from the census is also used in some instances). Each LSOA in England is then ranked from the most to the least deprived. Further details about the IMD are available from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/835115/IoD2019_Statistical_Release.pdf

Within the MCS, the LSOA in which children live is used to assign the IMD rank score. Data from the 2019 edition of the IMD are available from https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

Measurement of disadvantage / advantage

The IMD is often divided into quintiles, used to represent different socio-economic groups. We follow this approach, with the bottom IMD quintile used to measure 'disadvantage' and the top IMD quintile taken of measure 'advantage'.

In our analysis where we estimate the 'optimal' cut-points along the IMD scale, 34 percent of the population fall into the most disadvantaged group and 42 percent into the advantaged group.

Reference / Links

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/835115/IoD2019_Statistical_Release.pdf

https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/833951/IoD2019_Technical_Report.pdf

Income Deprivation Affecting Children (IDACI)

The IDACI index is a sub-scale of the Index of Multiple Deprivation (and, specifically, the income domain). It captures the proportion of 0-15-year-old children living in income deprived families with a Lower Super Output Area (LSOA). This is operationalised as families either in receipt of income support, income-based job-seekers allowance, income-based Employment and Support allowance, pension credit, universal credit, or in-receipt of working tax credit with an income below 60 percent of the national median. Each LSOA in England is then ranked from the most to the least deprived on this measure (I.e. from the largest to smallest proportion of families with children in the area being eligible for low-income benefits). Further details about the IDACI measure are available within the 2019 IMD technical report (see https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/833951/IoD2019_Technical_Report.pdf).

Within the MCS, the LSOA in which children live is used to assign the IDACI rank score. Data from the 2019 edition of the IDACI are available from https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

Measurement of disadvantage / advantage

IDACI is often divided into quintiles, used to represent different socio-economic groups. We follow this approach, with the bottom IDACI quintile used to measure 'disadvantage' and the top IDACI quintile taken of measure 'advantage'.

In our analysis where we estimate the 'optimal' cut-points along the IDACI scale, 37 percent of the population fall into the most disadvantaged group and 41 percent into the advantaged group.

Reference / Links

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/835115/IoD2019_Statistical_Release.pdf

https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/833951/IoD2019_Technical_Report.pdf

Free School Meals

Eligibility for Free School Meals (FSM) is a widely used proxy for low-income used in academic research, policy and practise in England. It is information routinely recorded within the National Pupil Database (NPD) as part of the regular school census. FSM are a means-tested benefit, though the criteria used to determine eligibility for FSM has changed over time (Hobbs and Vignoles 2010), with the current guidelines available from https://www.gov.uk/apply-free-school-meals. According to Joyce and Waters (2018), with the introduction of Universal Credit, "*the government has said that it will offer FSMs to families in receipt of UC who have annual net earnings (i.e. after income tax and employee National Insurance) of £7,400 or less*". Moreover, importantly, children are flagged as 'eligible' for FSM in the NPD only if they are both eligible for and claiming FSM (Hobbs and Vignoles 2010). This will mean that FSM, as measured in the NPD, will miss some low-income pupils (those who are eligible for this entitlement, but do not claim). Information on FSM is recorded each year that a child is enrolled in a state school. It is hence possible to use information from across multiple school years to calculate the proportion of time children spent at school which they eligible for FSM.

Information on FSM eligibility is available for MCS cohort members in England, via the link that has been made between the survey and the NPD. This is measured at the individual pupil level.

Measurement of disadvantage / advantage

When using a single year of FSM data, 'disadvantage' is simply defined as those eligible for FSM, with 'advantaged' defined as those who are not. Consequently, 17 percent of MCS cohort members are defined as disadvantaged, and 83 percent advantaged, when FSM information is drawn from a single year.

To create the 'permanent' FSM measure, we calculate the proportion of time children spent at school (between ages 5 and 16) that they were entitled to receive FSM. The distribution of this variable is provided in the table below. 'Disadvantage' is defined using this measure as those in the top quintile (i.e. the 20 percent of the population who were eligible for FSM for the greatest proportion of their time at school). The 'optimum' cut-point for disadvantage is also calculated for this measure, with 33 percent of the population then falling into the most disadvantaged group.

In contrast, 'advantaged' is defined as children who were never eligible for FSM during their time at school, while encompasses 67 percent of MCS cohort members.

Table B3. The distribution of percent of time at school MCS cohort members were eligible for FSM

% of time FSM eligible	%
Never FSM eligible	70%
0-10% of time	4%
11-20% of time	3%
21-30% of time	3%
31-40% of time	2%
41-50% of time	2%
51-60% of time	2%
61-70% of time	2%
71-80% of time	2%
81-90% of time	2%
91-100% of time	7%

References

Joyce, R. and Waters, T. 2018. Free school meals under universal credit. Accessed 31/03/2020 from https://www.ifs.org.uk/uploads/publications/bns/BN232.pdf

Graham Hobbs & Anna Vignoles (2010): Is children's free school meal 'eligibility' a good proxy for family income?, British Educational Research Journal, 36:4, 673-690.

Young Participation by Area Rate / POLAR

Polar is an indicator of university participation by local area. It is a key measure used in contextual admission in UK higher education, and is the preferred indicator of educational disadvantage by the Office for Students (the UK higher education regulator). It is a measure that captures how likely young people are to participate in higher education, depending upon the area that they live. Specifically, the 'young participation rate' is first calculated as the number of 18/19-year-olds from a given area who enter higher education and divide this by the total number of 18/19-year-olds who live in that area (i.e. the young participation rate is the proportion of young people within a given area who go to university). The area used is the Middle Super Output Area (MSOA), which contain around 7,500 individuals (across all ages). The 'Polar' classification is simply a categorised version of the youth participation rate, which divides this index into five quintiles.

Within the MCS, the MSOA in which children live is used to assign them Youth Participation Rate (YPR) scores and Polar quintiles. Data for the Youth Participation rate and Polar quintiles are available from https://www.officeforstudents.org.uk/data-and-analysis/young-participation-by-area/get-the-data/

Measurement of disadvantage / advantage

Following the use of Polar in university admissions, we define 'disadvantage' as the bottom quintile of the young participation rate (otherwise known as Polar Q1), and 'advantage' as the top quintile. Our analysis reveals that the 'optimal' cut point to define disadvantage occurs at the 54^{th} percentile (i.e. approximately 54 percent of the population falls into the YPR disadvantaged group) and at the 61^{st} percentile of the index to define advantage (i.e. approximately 39 percent of the population falls into the YPR advantaged group).

Links

https://www.officeforstudents.org.uk/data-and-analysis/young-participation-by-area/about-the-data/

TUNDRA

Tundra is an indicator of university participation by local area. It is an experimental measure of educational disadvantage developed by the Office for Students (the UK higher education regulator). It is a measure that captures how likely 16-year-old state school pupils are to participate in higher education by age 18/19. Specifically, it is the proportion of 16-year-olds from state schools who went on to higher education divided by the total number of 16-year-olds within a given area. The area used is the Middle Super Output Area (MSOA), which contain around 7,500 individuals (across all ages). The 'Tundra' classification is simply a categorised version of this university participation rate, which divides this index into five quintiles.

Within the MCS, the MSOA in which children live is used to assign them into Tundra quintiles. Tundra data are available from https://www.officeforstudents.org.uk/data-and-analysis/young-participation-by-area/get-the-data/

Measurement of disadvantage / advantage

Following the use of Tundra in university admissions, we define 'disadvantage' as the bottom Tundra quintile, and 'advantage' as the top quintile. Our analysis reveals that the 'optimal' cut point to define disadvantage occurs at the 49th percentile (i.e. approximately 49 percent of the population falls into the Tundra disadvantaged group) and at the 63rd percentile of the index to define advantage (i.e. approximately 37 percent of the population falls into the Tundra advantaged group).

Links

https://www.officeforstudents.org.uk/data-and-analysis/young-participation-by-area/about-the-data/

Carstairs Index

The Carstairs index was developed in 1991 (Carstairs and Morris 1991). It combines together four variables recorded in the census measured at the Lower Super Output Area (LSOA) level (LSOAs include around 1,500 individuals in approximately 650 households). The four variables that comprise the Carstairs index are (a) male unemployment; (b) lack of car ownership; (c) overcrowding and (d) low social class (see Brown, Allik, Dundas and Leyland 2014 for further details). Each of these variables is first standardised to mean zero and standard deviation one. These standardised scores are then summed together to create the final scale. The Carstairs index is therefore a continuous neighbourhood index of deprivation.

Within the MCS, the LSOA in which children live is used to assign them Carstairs Index scores. Data for the Carstairs index are available from http://reshare.ukdataservice.ac.uk/851497/

Measurement of disadvantage / advantage

We follow convention and define 'disadvantage' as the bottom quintile of the Carstairs index, and 'advantage' as the top quintile. Our analysis reveals that the 'optimal' cut point to define disadvantage occurs at the 45th percentile (i.e. approximately 45 percent of the population falls into the Carstairs disadvantaged group) and at the 58th percentile of the index to define advantage (i.e. approximately 42 percent of the population falls into the Carstairs advantaged group).

References

Brown, D.; Allik, M.; Dundas, R. and Leyland, A. 2014. Carstairs Scores for Scottish Postcode Sectors, Datazones & Output Areas from the 2011 Census.

Townsend Index

The Townsend index was developed in 1988 (Townsend, Phillimore and Beattie 1988). It combines together four variables recorded in the census measured at the Output Area (OA) level (OAs include approximately 125 households). The four variables that comprise the Townsend index are (a) unemployment; (b) lack of car ownership; (c) overcrowding and (d) non-home ownership (see Yousaf and Bonsall 2017 for further details). To create the Townsend Index, the unemployment and overcrowding indicators were first log-transformed. Each of the four variables were then standardised to mean zero and standard deviation one. These standardised scores are then summed together to create the final scale. The Townsend index is therefore a continuous neighbourhood index of deprivation.

Within the MCS, the output area in which children live is used to assign them Townsend Index scores. Data for the Carstairs index are available from https://www.statistics.digitalresources.jisc.ac.uk/dataset/2011-uk-townsend-deprivation-scores

Measurement of disadvantage / advantage

We follow convention and define 'disadvantage' as the bottom quintile of the Townsend index, and 'advantage' as the top quintile. Our analysis reveals that the 'optimal' cut point to define disadvantage occurs at the 30th percentile (i.e. approximately 30 percent of the population falls into the Townsend disadvantaged group) and at the 58th percentile of the index to define advantage (i.e. approximately 42 percent of the population falls into the Townsend advantaged group).

References

Yousaf, S. and Bonsall, A. 2017. UK Townsend Deprivation Scores from the 2011 census data. Accessed 31/03/2020 from 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

IFS socio-economic status index

The Institute for Fiscal Studies (IFS) measure of socio-economic status was first developed in a paper by Chowdry et al (2010). It has since been used in a relatively small number of academic papers (e.g. Burgess, Crawford and Macmillan 2018; Campbell et al 2019). It combines together information from a number of the proxies detailed above, in particular FSM eligibility, the IMD, Acorn and a number of census variables (similar to the Carstairs and Townsend indices). In its original carnation, which we follow in this paper, the IFS socio-economic status index included the following information:

- Eligibility for Free School Meals (FSM) at age 16
- Index of Multiple Deprivation score (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)

These variables are included within a principal components analysis, with the first component (which explains around 60 percent of the variance) forming the index. This scale is then standardised to mean zero and standard deviation one. The IFS socio-economic scale is therefore a continuous index of socio-economic background.

Measurement of disadvantage / advantage

We define 'disadvantage' as the bottom quintile of the IFS socio-economic status index, and 'advantage' as the top quintile. Our analysis reveals that the 'optimal' cut point to define disadvantage occurs at the 40th percentile (i.e. approximately 40 percent of the population falls into the IFS scale disadvantaged group) and at the 64th percentile of the index to define advantage (i.e. approximately 36 percent of the population falls into the IFS scale advantaged group).

References

Burgess, S.; Crawford, C.; and Macmillan, L. 2018. Access to grammar schools by socio-economic status. Accessed 31/03/2020 from https://discovery.ucl.ac.uk/id/eprint/10053242/1/Burgess,%20Crawford%20and%20Macmillan_GrammarandSES_Revised.pdf

Campbell, S.; Macmillan, L,; Murphy, R. and Wyness, G. 2019. Inequalities in Student to Course Match: Evidence from Linked Administrative Data. Accessed 31/03/2020 from http://cep.lse.ac.uk/pubs/download/dp1647.pdf

Chowdry, H.; Crawford, C.; Dearden, L.; Goodman, A. and Vignoles, A. 2010. Widening participation in higher education: analysis using linked administrative data. IFS working paper W10/04. Accessed 31/03/2020 from https://www.ifs.org.uk/wps/wp1004.pdf

Appendix B. Alternative estimates using a multidimensional measure of socio-economic status

In the main body of the text, we investigate how well each of the measures acts as a proxy for permanent family income. There are, however, alternative measures of socio-economic background, including parental education, social class (based upon parental occupation) or a combination of multiple indicators. In this paper, we provide a set of alternative estimates, where we consider how well each measure proxies a multi-dimensional measure of socio-economic status. This multidimensional measure is a combination of the following socio-economic indicators (measured at age 14 unless otherwise stated)[6]:

- Maternal education (highest academic qualification held)
- Paternal education (highest academic qualification held)
- Maternal social class (NS-SEC of current or most recent occupation held)
- Paternal social class (NS-SEC of current or most recent occupation held)
- Permanent family income quintile

An item-response model (graded response model for ordinal items) is used to combine these variables into an overall socio-economic status scale. A histogram illustrating the distribution of this scale is provided in Figure B1 below. The correlation between the final scale and each variable used in its construction are as follows:

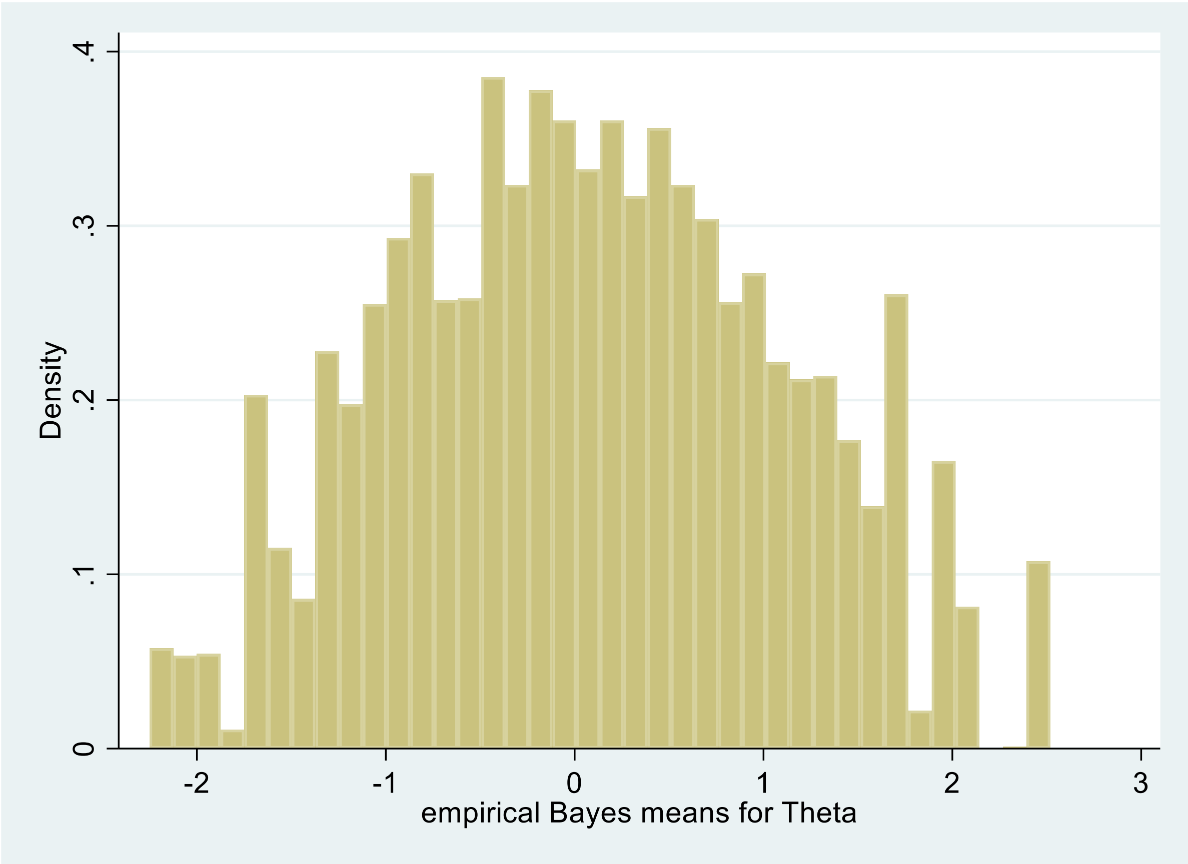
- 0.79 with maternal education
- 0.77 with paternal education
- 0.72 with maternal social class
- 0.75 with paternal social class
- 0.86 with permanent family income

Using this scale, 'low' socio-economic status is defined as those cohort members who fall into the bottom quintile of the distribution (analogous to the bottom quintile of the permanent income distribution being used to define income poverty in the main body of the paper). Similarly, 'high' socio-economic status is defined as those who fall into the top quintile of the distribution (analogous to the top quintile of the permanent income scale being used to define income affluence).

[6] If data are not available at age 14, information on the socio-economic status indicator is carried forward from the previous wave.

All key results presented in the tables and figures are reproduced using this multidimensional socio-economic scale as the outcome variable of interest (rather than permanent family income). These are presented below.

Appendix Figure B1. Distribution of the multi-dimensional socio-economic status scale



Appendix Table B1. The correlation between different proxy measures and permanent family income

Measure	Type	Correlation with poverty	Correlation with permanent income	Correlation with affluence
IMD	Age 14	0.51	0.49	0.50
	Permanent	0.54	0.53	0.58
FSM	Age 14	0.65	0.41	-
	Permanent	0.69	0.53	-
IDACI	Age 14	0.50	0.50	0.49
	Permanent	0.55	0.54	0.57
YPR/POLAR	Age 14	0.28	0.37	0.48
	Permanent	0.29	0.38	0.51
IFS	Age 14	0.56	0.58	0.62
	Permanent	0.58	0.62	0.66
ACORN	Age 14	0.61	0.56	0.63
	Permanent	0.66	0.61	0.68
Carstairs index	Age 14	0.53	0.51	0.54
	Permanent	0.55	0.54	0.58
Townsend index	Age 14	0.53	0.51	0.44
	Permanent	0.57	0.55	0.51
TUNDRA	Age 14	0.25	0.30	0.37
	Permanent	0.22	0.31	0.40
OAC	Age 14	0.52	0.45	0.53
	Permanent	0.49	0.51	0.39
Income age 14	Age 14	0.58	0.59	0.78

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

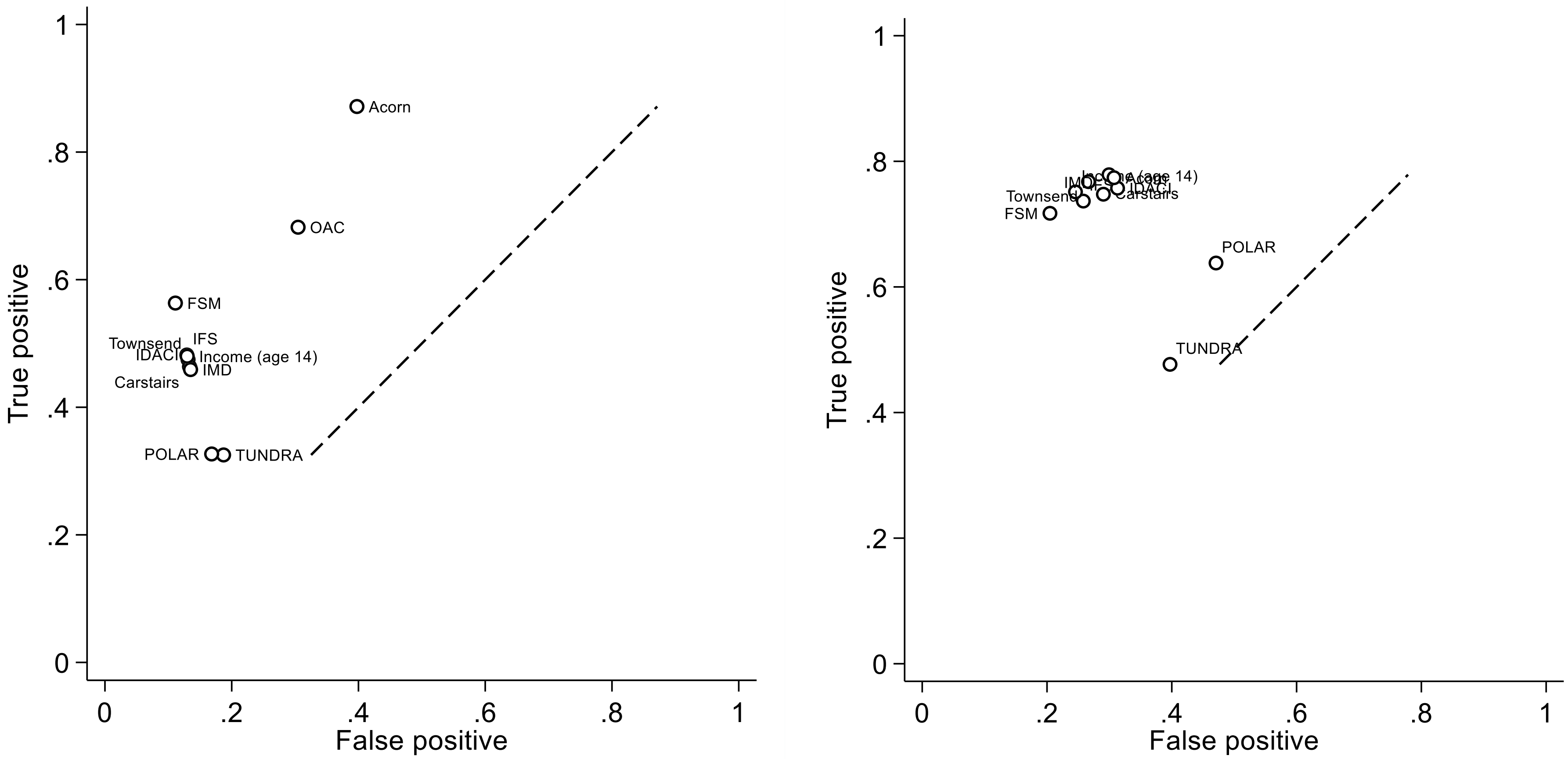
Appendix Table B2. Bias in the proxies as a measure of permanent low-income

Measure	Type	London	Ethnic minority	Single parent	Renter	Male	Young mother	Young father
IMD	Age 14	8%	12%	18%	25%	-1%	21%	7%
	Permanent	9%	11%	17%	25%	-1%	20%	7%
FSM	Age 14	7%	14%	11%	19%	-1%	19%	7%
	Permanent	3%	11%	8%	15%	-1%	16%	7%
IDACI	Age 14	4%	12%	18%	25%	-1%	21%	8%
	Permanent	3%	12%	17%	24%	-1%	20%	7%
YPR/POLAR	Age 14	12%	20%	20%	29%	-1%	24%	9%
	Permanent	12%	21%	21%	29%	-1%	24%	9%
IFS	Age 14	7%	13%	17%	23%	-1%	20%	7%
	Permanent	8%	13%	17%	23%	-1%	19%	7%
ACORN	Age 14	9%	12%	15%	22%	-1%	17%	4%
	Permanent	9%	13%	15%	21%	-1%	15%	2%
Carstairs index	Age 14	3%	8%	18%	25%	-1%	21%	7%
	Permanent	2%	6%	18%	25%	-1%	21%	7%
Townsend index	Age 14	-1%	7%	17%	24%	-1%	22%	8%
	Permanent	-3%	5%	17%	23%	-1%	20%	7%
TUNDRA	Age 14	12%	20%	21%	29%	-1%	25%	9%
	Permanent	12%	22%	21%	29%	-2%	22%	7%
OAC	Age 14	9%	13%	18%	25%	-1%	21%	7%
	Permanent	9%	12%	17%	24%	-1%	19%	5%
Income age 14	Age 14	7%	13%	16%	26%	-1%	23%	11%

Notes: Figures indicate how much more likely the group be in the bottom socio-economic scale quintile, conditional upon the proxy measure. For instance, Londoners are around 12 percentage points more likely to be in the bottom socio-economic quintile than those living elsewhere in England, conditional upon the POLAR 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).

Appendix Figure B2. True-positive and false-positive rates for detecting low-SES using different proxy measures

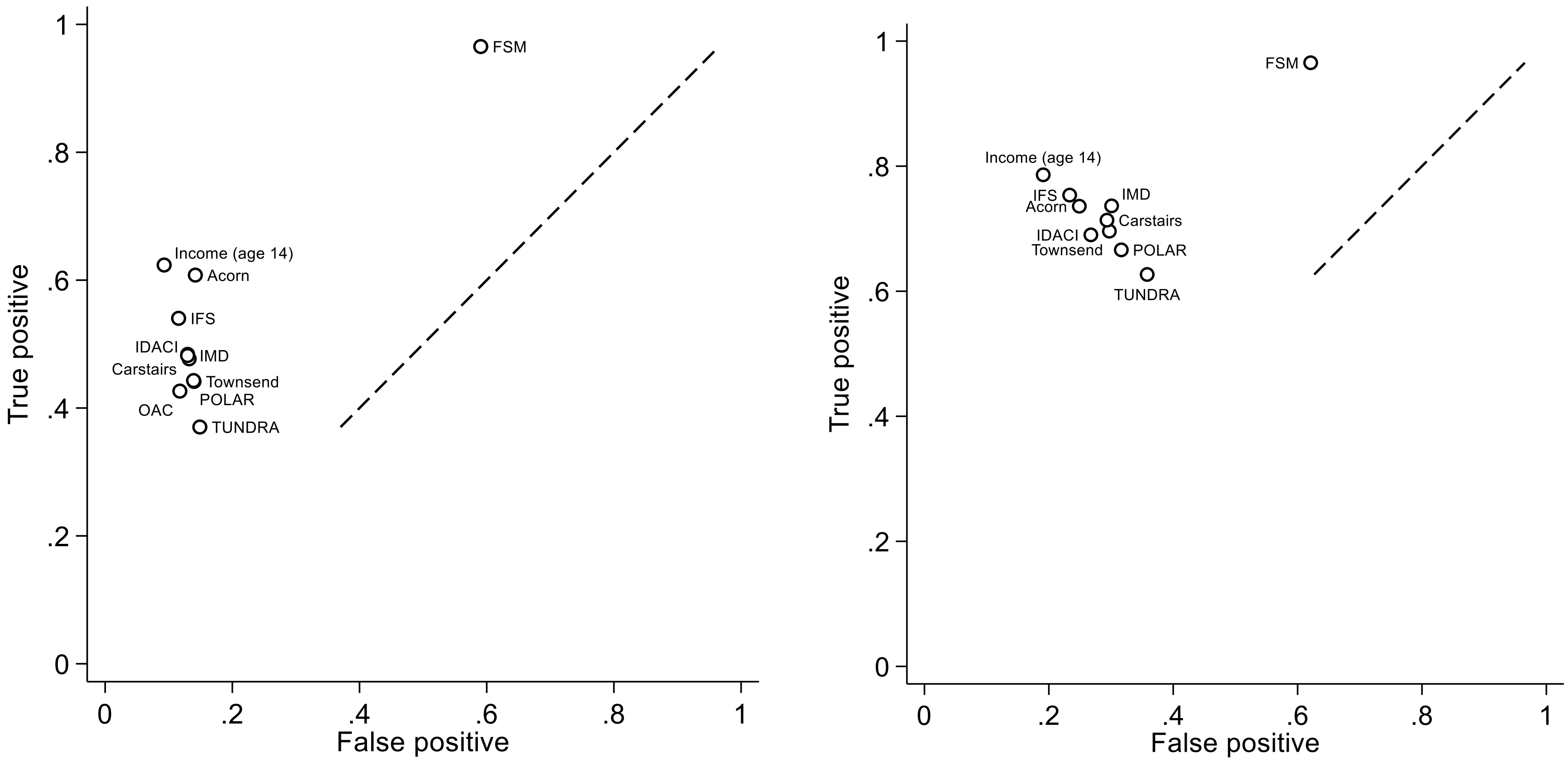
(a) Conventional cut-point (b) Optimal-cut-point



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.

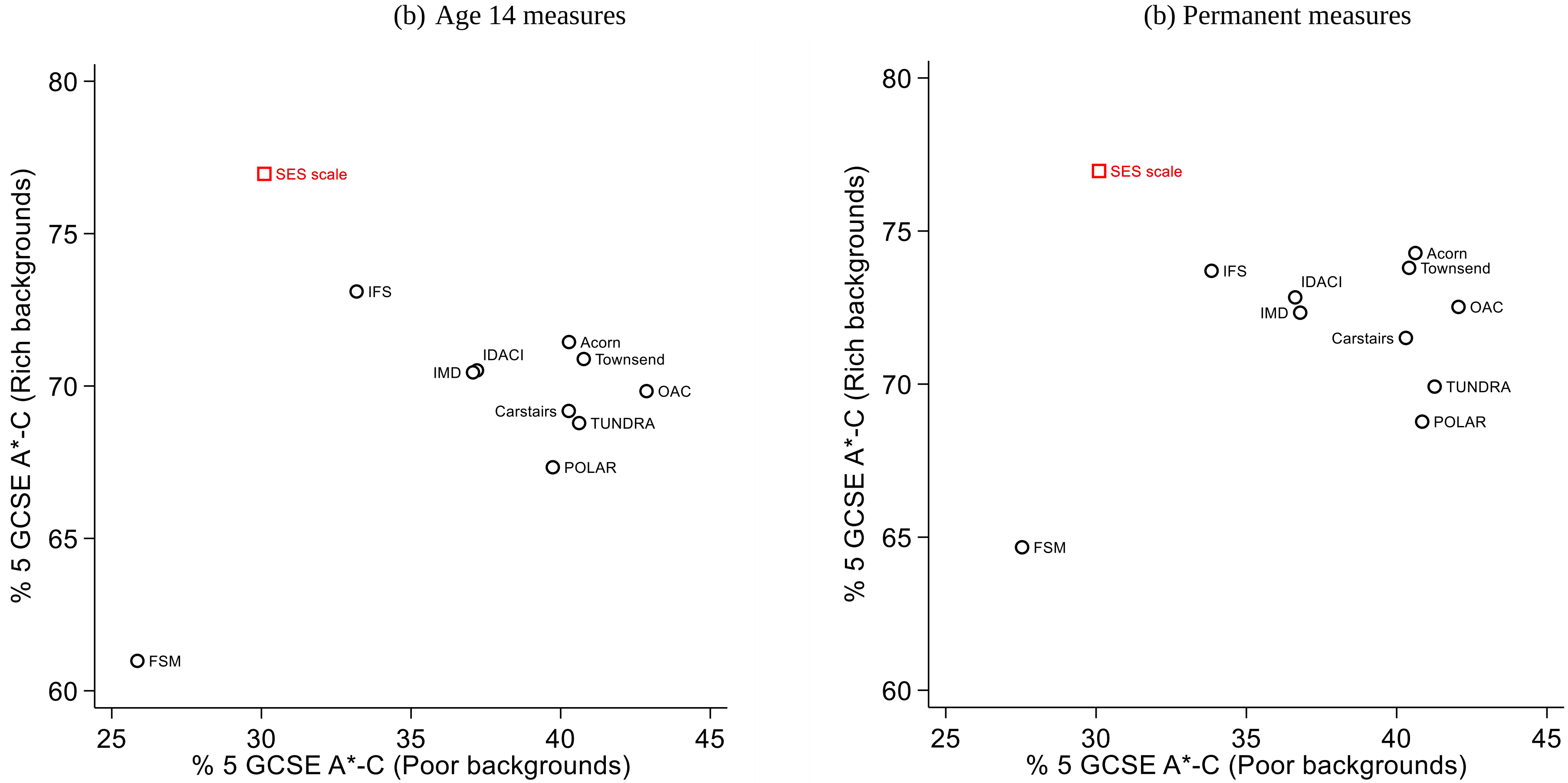
Appendix Figure B3. True-positive and false-positive rates for detecting high-SES using different proxy measures

(b) Conventional cut-point (b) Optimal-cut-point



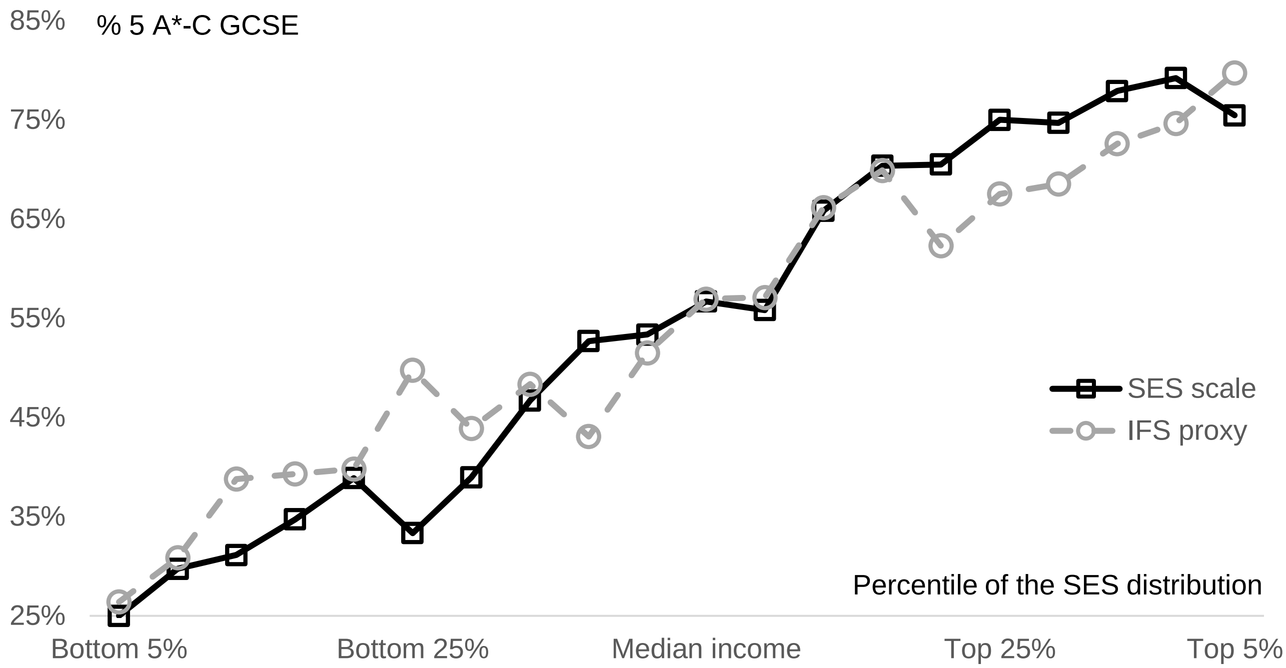
Notes: True-positive rate is presented on the vertical axis; this captures the percent of high-income families correctly identified by the proxy. The false-positive rate is plotted along the horizontal axis, capturing the percent of families the proxy identifies as coming from a high-income background 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.

Appendix Figure B4. How well do the proxy measures capture socio-economic gaps in academic achievement? (Conventional cut-point).



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.

Appendix Figure B5. The percent of children achieving five A*-C grades by position in the SES distribution. IFS measure versus the socio-economic scale.



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

Appendix C. Bias in each proxy as a measure of permanent income and permanent income affluence

(a) Permanent income

Measure	Type	London	Ethnic minority	Single parent	Renter	Male	Young mother	Young father
IMD	Age 14	10	-67	-176	-233	6	-161	-84
	Permanent	7	-50	-166	-216	7	-139	-70
FSM	Age 14	13	-115	-168	-265	6	-195	-116
	Permanent	39	-89	-126	-217	6	-157	-106
IDACI	Age 14	57	-67	-172	-227	7	-157	-85
	Permanent	65	-52	-160	-205	7	-137	-72
YPR/POLAR	Age 14	-108	-182	-199	-266	4	-184	-96
	Permanent	-130	-186	-195	-259	6	-176	-86
IFS	Age 14	11	-71	-145	-182	6	-119	-60
	Permanent	11	-54	-133	-156	7	-97	-47
ACORN	Age 14	22	-85	-137	-173	8	-125	-61
	Permanent	21	-76	-128	-142	12	-92	-36
Carstairs index	Age 14	87	-9	-173	-229	6	-161	-84
	Permanent	100	23	-166	-214	8	-145	-73
Townsend index	Age 14	141	-7	-165	-222	8	-164	-85
	Permanent	159	23	-155	-199	10	-138	-70
TUNDRA	Age 14	-108	-210	-207	-279	7	-201	-109
	Permanent	-129	-224	-207	-276	8	-196	-102
OAC	Age 14	-9	-99	-188	-248	8	-181	-97
	Permanent	-9	-79	-177	-227	7	-159	-80
Income age 14	Age 14	-13	-71	-47	-147	2	-137	-77

(b) Income affluence

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%

Appendix D. The percent of cohort members in the ‘disadvantaged’ and ‘advantaged’ group for each proxy measure

	Income poverty		**Income affluence**	
	Conventional	**Optimal**	**Conventional**	**Optimal**
Acorn	49%	41%	23%	35%
OAC	38%	N/a	18%	N/a
IMD	20%	34%	20%	42%
IDACI	20%	37%	20%	41%
FSM age 14	17%	N/a	83%	N/a
Years of FSM	20%	33%	67%	67%
YPR/POLAR	20%	54%	20%	39%
Tundra	21%	49%	20%	37%
Carstairs	20%	45%	20%	42%
Townsend	20%	30%	20%	42%
IFS	20%	40%	20%	36%
Single income	20%	34%	20%	27%
Permanent income	**20%**		**20%**	