THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Using Linked Longitudinal Administrative Data to Identify Social Disadvantage

OPEN ACCESS

ORIGINAL RESEARCH

Using Linked Longitudinal Administrative Data to Identify Social Disadvantage

Serena Pattaro[1] (ID) · **Nick Bailey**[1] (ID) · **Chris Dibben**[2]

Abstract

Administrative data are widely used to construct indicators of social disadvantage, such as Free School Meals eligibility and Indices of Multiple Deprivation, for policy purposes. For research these indicators are often a compromise between accuracy and simplicity, because they rely on cross-sectional data. The growing availability of longitudinal administrative data may aid construction of more accurate indicators for research. To illustrate this potential, we use administrative data on welfare benefits from DWP's National Benefits Database and annual earnings from employment from HMRC's P14/P60 data to reconstruct individual labour market histories over a 5-year period. These administrative datasets were linked to survey data from the Poverty and Social Exclusion UK 2012. Results from descriptive and logistic regression analyses show that longitudinal measures correlate highly with survey responses on the same topic and are stronger predictors of poverty risks than measures based on cross-sectional data. These results suggest that longitudinal administrative measures would have potentially wide-ranging applications in policy as well as poverty research.

Keywords Administrative data · Administrative and survey data comparison · Earnings · Poverty and social disadvantage · Unemployment · Welfare benefits

1 Introduction

Administrative data, particularly those from the welfare benefits and tax systems, are employed in the UK to construct indicators of social disadvantage for use across a range of policy domains. The ready availability of these indicators as well as their policy salience means that they are very attractive to researchers interested in social inequalities. However, their origin in policy and practice means they are often highly compromised in their

---

✉ Serena Pattaro
   Serena.Pattaro@glasgow.ac.uk

1   Scottish Centre for Administrative Data Research, School of Social and Political Sciences, University of Glasgow, Urban Big Data Centre, 7 Lilybank Gardens, Glasgow G12 8RZ, UK

2   Scottish Centre for Administrative Data Research, Institute of Geography and the Lived Environment, School of Geosciences, University of Edinburgh, Drummond Street, Edinburgh EH8 9XP, UK

   🍎 Springer

design, trading off accuracy for simplicity or ease of implementation. They frequently rely on simple cross-sectional indicators derived from a single source. Classic examples are eligibility for Free School Meals, based on whether a family currently receives particular means-tested welfare benefits, and Indices of Multiple Deprivation, which combine indicators from multiple domains, including income and employment derived from welfare benefits and tax credits.

With the growing availability of administrative data for research purposes, researchers have the opportunity to link together multiple sources of data over multiple years, and hence to design and implement their own, more sophisticated measures of social disadvantage, free from the operational constraints facing policy makers. They have the opportunity to exploit longitudinal information on individuals rather than rely solely on cross-sectional data, and to combine information from different domains to identify social disadvantage with greater accuracy. The aim of this paper is therefore to explore the potential to identify social disadvantage using linked longitudinal administrative data. We assess the approach through the analysis of longitudinal administrative data on welfare benefits and earnings which have been linked to a major UK survey of poverty. We compare how well longitudinal measures of benefit receipt and earnings predict current poverty risks, and compare these with survey measures on the same topic as well as cross-sectional measures from administrative data.

1.1 Education Policy and Free School Meals

One area which has made extensive use of indicators of social disadvantage is education, particularly since the introduction of the Free School Meals (FSM) policy in 1980. In that year, the duty on local education authorities to provide school meals for all children (at a charge) was replaced with a duty to provide free meals for the children of low-income families (Gustafsson 2002; Evans and Harper 2009). This shift from universalist to means-tested provision brought with it a requirement to identify those eligible for FSM. While the precise definition has shifted over time, the principles have remained the same. As the Department for Education states in its discussion of the latest revisions, the indicator of eligibility is aiming to target needs effectively but also to provide "a clear and simple system that is realistic for schools and local authorities to deliver" (Department for Education 2018; p. 10). In other words, the indicator is necessarily a pragmatic compromise.

Eligibility for FSM is determined by whether a family is in receipt of particular means-tested welfare benefits. As the welfare system changes, the list of benefits which lead to eligibility gets revised, and this has sometimes led to rather arbitrary shifts in entitlement. At the outset in 1980, for example, the 'passport' to FSM was Supplementary Benefit, which was available to those on low incomes regardless of employment status. With a major change in the welfare benefits system in 1988, in-work and out-of-work benefits were split with eligibility for FSM initially limited to those receiving the new out-of-work benefit, Income Support (Gustafsson 2002). With the recent introduction of a new single benefit, Universal Credit, covering those in and out of work, FSM eligibility is available to both but, as the benefit is claimed by quite a wide range of groups, eligibility has been restricted to those who also pass a 'low earnings' test (Department for Education 2018). Such changes in eligibility criteria may be necessary for pragmatic reasons but they have nothing to do with changes in social need or disadvantage, raising the question whether they can be effectively used to accurately capture social disadvantage for research purposes.

The use of the FSM eligibility indicator has been extended to other areas of educational policy and practice over time. Since 2011, for example, it has been the basis for allocating additional resources to English schools (the 'Pupil Premium') with the aim of reducing the educational attainment gap between children from disadvantaged backgrounds and their better-off peers (House of Commons Education Committee 2014). Since this indicator is implemented at the level of schools, the requirement for simplicity can be relaxed and there has therefore been a move to use more longitudinal information. Additional funding is determined by the number of children who were eligible for FSMs at any point in the previous 6 years (the so-called 'Ever-6 FSM' indicator) (Department for Education 2017).

From its origins as a device for policy, FSM eligibility has been taken up as an indicator of social disadvantage in a wide range of research on educational attainment and interventions, driven primarily by the fact that it is readily available (Kounali et al. 2008). As the then-Minister for Education said in 2005, "We have no data on the social class of the parents of children in school … so we proxy social class by whether or not the pupil is in receipt of FSM" (quoted in Kounali et al. 2008; p. 2). This has prompted researchers to examine whether FSM receipt is an accurate and reliable proxy indicator of pupils' socio-economic circumstances. Eligibility for FSM is preferred over actual claims for FSM as many families do not take up the service for reasons of stigma, as well as cultural, dietary or other factors (Gorard 2012; Lord et al. 2013). Even so, FSM eligibility fails to capture many children living in low-income households who we might expect to face educational disadvantages (Kounali et al. 2008; Hobbs and Vignoles 2010). Compared to alternative cross-sectional indicators of social disadvantage, including household income and area deprivation measures, FSM eligibility displayed only a marginally lower predictive power of pupils' educational attainment (Ilie et al. 2017). Nevertheless, the question remains whether better measures could be constructed for research purposes from linked longitudinal administrative data.

1.2 Neighbourhood Policy and Indices of Multiple Deprivation

A second area where administrative data indicators have come to be widely used is in relation to neighbourhood policy and analysis. In the UK, there is a long history of neighbourhoods with high concentrations of social disadvantage and associated problems being identified through composite indices which combine a large number of indicators. Neighbourhoods regarded as deprived can then be targeted for additional resources, services or interventions through a variety of mechanisms known broadly as 'area based initiatives'. There are various arguments to justify such approaches although they are also seen as controversial by some; Manley et al. (2013) provide a review.

For a long time, deprivation indices relied solely on Census data, limiting the frequency of updates (e.g. Holtermann 1975; Jarman 1983; Department of the Environment 1994). With the New Labour government's strong interest in the spatial dimensions of poverty and social exclusion (Social Exclusion Unit 2001; 2004), it began to develop indices which were based primarily on administrative data sources and which could therefore be updated more frequently (Department of the Environment Transport and the Regions 2000). These were further developed into the current Indices of Multiple Deprivation (IMDs), a model which came to be used across the UK albeit with minor local variations (Noble et al. 2006). In contrast to the narrow focus on welfare benefits receipt for FSM eligibility, these indices are designed to cover the multiple domains of disadvantage which are seen as part of the concept of neighbourhood deprivation (Bailey et al. 2003). These include attributes of

residents (e.g. poverty, unemployment and poor health), as well as aspects of the physical context (e.g. poor housing or physical environment) and of the social environment (e.g. crime). Within the current indices, measures of income and employment deprivation are the dominant factors, and both are derived from cross-sectional data on welfare benefits and tax credits.

Indicators of neighbourhood deprivation are easily attached to a wide range of administrative, Census and survey data making them widely available. As a result, they have been taken up by researchers concerned with social inequalities in many different fields including education and health (e.g. Barnes et al. 2006; Boyle et al. 2004; Norman et al. 2011; Strand 2011). For education research, for example, Kounali et al. (2008) note that the main alternative to the FSM eligibility indicator discussed above is the IMD.

The IMD approach is subject to a number of criticisms. One of the most common, from Holtermann (1975) onwards, is that the majority of those who are 'deprived' do not live in 'deprived neighbourhoods'; some would argue that this misses the point since a deprived neighbourhood is more than simply the sum of its residents. Another criticism is that the focus on deprived places obscures the extent to which the deprived population resident within them is highly dynamic, with individuals moving in and out of poverty, and moving in and out of the neighbourhood (Bailey et al. 2013; Gambaro et al. 2016; Fransham 2017). Aggregate levels of deprivation may mask variations between places in the extent to which they are home to individuals facing long-term poverty or a fast-changing group of people experiencing shorter-term spells.

1.3 The Potential of Linked Longitudinal Administrative Data

There are two motivations for considering the use of indicators derived from linked longitudinal administrative data. The first is that the concept of social disadvantage, like that of poverty, is inherently a longitudinal one. As Townsend (1979; 1987) argued, poverty arises from the lack of command of resources over time. Cross-sectional measures based on current income, such as those in the annual reporting of UK poverty levels (Department for Work and Pensions 2018), can only provide an approximation. Similarly, social disadvantage implies a prolonged situation of life with low levels of resources, restricting material consumption as well as social engagement. Yet, as Kounali et al. (2008) show, measures based on current benefit status such as eligibility for FSM are very volatile.

The second motivation is that, as in many other countries, researchers in the UK are finding increasing access to administrative data and hence the opportunity to get beyond the restrictions of already-available measures and to design their own. Following high-level government commitment to realise the potential value of administrative data for policy-relevant research (Cabinet Office 2012) and consultations by the Administrative Data Taskforce (2012), the UK established the Administrative Data Research Network (ADRN), a service promoting access to linked administrative data for academic and non-academic researchers (http://www.adrn.ac.uk). While progress in securing access to data held by UK government departments has not been as rapid as hoped (UK Statistics Authority 2017), there is a clear potential for scientific and policy-related research. Investments in this area have been renewed through Administrative Data Research UK (https://www.adruk.org/) which, working closely with the Office for National Statistics (ONS) and key policy makers and stakeholders, aims to explore important cross-cutting research questions using and promoting linked administrative data.

As Smith et al. (2004) noted, linked administrative records are potentially well suited for longitudinal analysis. Compared with conventional survey data, administrative records may contribute to lowering the costs of data collection, in terms of financial and time resources (e.g. by requiring fewer questions during interviews) and potentially provide a wider coverage of the target population. Administrative records may also help to reduce the burden on respondents and they are less affected by declining response rates, attrition and loss to follow up, compared to longitudinal survey data. Linking administrative records over time can further enrich survey data, by providing updated longitudinal information before and between surveys, and help to construct consistent measures across different surveys.

However, several challenges arise from administrative data, with important implications for scientific research. Compared to survey data providing detailed information on a wider range of covariates, administrative data often exhibit a more limited coverage of information which can only be used to construct proxy indicators in social science research. Moreover, issues of comparability over time may arise from discontinuities in the definition of the information collected (e.g. eligibility criteria) due to changes in the government policy legislation. There are additional statistical and methodological issues associated with administrative data (for a review, see Hand (2018)). For example, administrative data may be the outcome of selection processes and, as a result, predictions are often generated from a 'sub-sample' of records which may be not representative of the whole population of reference. In other words, coverage of administrative data may be limited to individuals captured through their 'successful' interaction with the administrative process (i.e. by providing complete or non-missing information). Consent bias may also be introduced when survey respondents are asked for permission to link administrative records to information collected in surveys. Individuals who provide consent to administrative data linkage may differ systematically from individuals who refuse consent and this can affect statistical inferences drawn from linked administrative data (Kho et al. 2009; Moore et al. 2018; Sakshaug et al. 2012; Sala et al. 2014).

In addition, there is the uncertainty inherent in the linkage process which introduces new sources of measurement error (e.g. Bohensky 2016; Gilbert et al. 2017; Harron et al. 2017a; Zhang 2012), which also apply to the case of the construction of longitudinal records. There are also important questions about the validity of any proposed measures. This can be assessed in part by direct comparison with survey recall questions or even longitudinal survey responses where these are available. Of course, it is not clear that survey data is the 'gold standard', especially given the stigma attached to a status such as unemployment (Paugam and Russell 2000). Alternatively it can be assessed by the ability of measures to predict current risks of poverty.

In this paper, we take a step back from the two introductory examples concerning FSM and IMD. We focus on labour market participation captured via welfare benefits receipt as the core components of these measures. We explore the potential of these components to identify social disadvantage using linked longitudinal administrative data through an empirical study. We construct a dataset of linked longitudinal administrative data from welfare benefits and tax systems, linking this in turn to data from a large national survey of poverty and social exclusion. We use the survey data to compare and validate measures of disadvantage derived from benefits and earnings histories. We explore the advantages of moving from cross-sectional to longitudinal measures from a single source, and the advantages of moving from a single domain (welfare benefits) to two domains (adding earnings). In particular, we address the following four research questions, of which research questions (a) and (b) can be considered as preliminary, preparing the ground for research questions (c) and (d) which lie at the core of this paper:

(a) How do labour market participation measures reconstructed from welfare benefits data compare with those derived from survey recall questions on the same topic?
(b) Do measures of current and past labour market participation derived from welfare benefits data predict poverty risks as well as or better than those based on survey recall questions?
(c) Do longitudinal measures derived from welfare benefits data predict current poverty risks better than measures based on current benefits receipt?
(d) Does the addition of longitudinal earnings measures to models including longitudinal benefits measures improve the prediction of poverty risks?

The remainder of the paper is organised as follows. Section 2 describes the data sources which were used. We explore the consequences for our available sample of the need for survey respondents' consent to administrative data linkage. We also discuss the validation and modelling approaches applied to the linked survey-administrative data. Section 3 presents the results from internal validation followed by the results from the descriptive analysis and logistic regression models. The paper concludes with a summary of the main results and a discussion on limitations and further research directions.

2 Data and Methods

2.1 Data Sources Used for Linkage

The analysis in this paper is based on a unique combination of survey data from two national surveys of incomes and poverty, linked to data from national administrative systems on welfare benefits and employment. The *Poverty and Social Exclusion in the UK* (PSE-UK) survey is a cross-sectional household survey conducted in 2012. It comprises a sample of respondents who had previously taken part in the *Family Resources Survey* (FRS) 2010/11 and given permission to be re-contacted. The FRS is an annual cross-sectional survey used to derive official statistics on income and poverty (Department for Work and Pensions 2018). The PSE-UK survey added to that from the FRS, in particular through a wider range of measures of poverty and deprivation (Bailey and Bramley 2018). In addition to permission for re-contact, FRS respondents in Britain are also asked for consent for various administrative data to be linked to their survey responses. As permission for linkage is not sought in Northern Ireland, that region is excluded here.

Within Britain, the FRS was based on a stratified clustered probability sample design with postcode sectors for the primary sampling units and households selected at random within these units. Surveys comprise face-to-face interviews with each resident aged 16 or over, using computer-assisted personal interviewing (CAPI). The PSE-UK sampled at random from FRS respondents who had given permission for re-contact, so retaining the latter survey's complex design. The FRS covered approximately 25000 households in UK, with a response rate of 59%. PSE-UK interviews were achieved with 4205 households in Britain (63% response rate). Within these households, full or partial interviews were achieved with 9786 individual respondents.

Data from the FRS and PSE-UK were linked to administrative data from the DWP's *Work and Pensions Longitudinal Study* (WPLS). This combines welfare benefits records held by DWP with earnings and employment information from Her Majesty's Revenue and Customs (HMRC) tax records. DWP and HMRC's datasets can be linked using unique

person identifiers, usually the National Insurance Number. Probabilistic linkage methods based on a set of identifying variables such as name, gender, date of birth and residential address were used by DWP to link the administrative data to the survey data (communication with DWP). Only anonymised, linked data are released to the research team. For the purposes of the current work, we assume that linkage errors are small and do not have important consequences for model estimates. We present descriptive results from the investigation of consent to administrative data linkage and the internal validation of measures derived from survey and administrative data sources to provide additional reassurance (see subsequent Sects. 2.2. and 3.1).

Three WPLS data sources were linked although only two are used in this analysis. The first was DWP's National Benefits Database (NBD) which provided data on receipt of two out-of-work benefits, Jobseeker's Allowance (JSA) and Income Support (IS). JSA is paid to various groups of working-age adults who are unemployed and actively seeking work. Some qualify on the basis of recent employment history (i.e. their 'contributions' through the tax system) while others qualify on the basis of low income. Additional eligibility criteria include low savings and partner's employment status. IS is paid to various groups of working-age adults on low income who are ineligible for JSA or for long-term sickness benefits, usually lone parents. Both benefits are in the process of being replaced by a new single benefit, Universal Credit, currently being rolled out across Great Britain, but this does not affect the period examined here.

The second WPLS source was HMRC's data on annual earnings, derived from P14/P60 tax return forms. These are issued by every employer at the end of each tax year to detail employees' taxable income in that employment across the whole year; the P60 is the part of the P14 which is given to the employee. Respondents may therefore have multiple returns in any year. HMRC earnings and employment records cover only employees and not self-employed individuals. Furthermore, tax records do not cover all employees as employers are not required to issue tax return forms if their employees' earnings are below the income tax threshold.

The third, unused dataset was HMRC's employment spell data, derived from P45/P46 tax return forms. The P45 is produced by employers when an individual leaves employment and details earnings and tax paid in the current tax year. This is given to the employees next employer to ensure accurate calculation of tax. A P46 is generated by the employer where a new employee does not have a P45 because they have no previous employment in the tax year, they have lost their P45 or they are starting an additional job alongside an existing one. We found the quality of the data from these forms was poor with many missing start and end dates, and large numbers of apparently duplicate records as documented by other studies (e.g. Barnes et al. 2011).

We linked administrative records covering the period of approximately 10 years up to and including the 2012/2013 tax year, the year in which PSE-UK survey was conducted. In this study, we focus on the last 5 years before the survey as this is the reference period covered by the survey recall questions that we use for validation purposes. It is also the period where earnings and unemployment are most likely to impact on current risks of poverty. At the time of this study, we had access to only a limited range of administrative data. Most obviously, we did not have data on benefits for people unable to work because of sickness or disability, or those with a caring role. What we have is a relatively limited extract which we use for the purposes of demonstrating our approach through one particular application: the use of benefits for unemployment and of tax records for employment in the effort to predict current poverty risks. Individual researchers will want to use different sets of data to identify forms of social need or disadvantage appropriate to their particular context. It
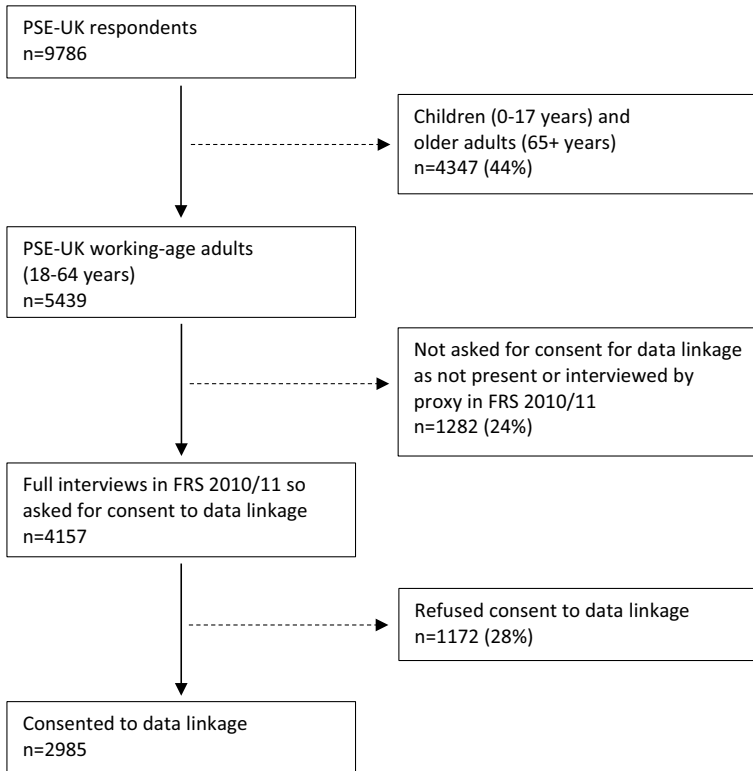
Fig. 1 Selection of the PSE-UK study sample

is important to be aware of these limitations when interpreting the results derived from the exercise presented in this paper.

2.2 Consent to Administrative Data Linkage and Sample Selection

Permission for administrative data linkage was only sought from respondents who participated fully in the original FRS interview rather than having their details included through a proxy interview with another household member. Of those completing a full interview in the FRS, 64% agreed to administrative data linkage. A larger proportion (83%) consented to being re-contacted (correspondence with DWP). Within this group, which forms the basis for the PSE-UK sample, consent for administrative data linkage was slightly higher (72%).

Figure 1 outlines the selection process that was followed to obtain the study sample for the current paper. Starting with all PSE-UK respondents (those present in the household at the PSE-UK survey date), we excluded children and adults 65 years or older to leave 5439 working-age adults. We then removed those not present at the time of the FRS interview or included only by a proxy interview because they were never asked for permission for data linkage. Of the remaining PSE-UK respondents who had provided a full FRS interview, 28% refused consent for linkage and were therefore excluded from the analyses in the core of this paper. That leaves 72% who consented to administrative data linkage. This is the

same consent rate as found among FRS respondents overall although in this case, the figure is limited to working-age adults only.

Table 1 shows the composition of the PSE-UK working-age sample (n = 5439) and provides a comparison between three groups. These include those who: (a) consented to data linkage; (b) refused consent for data linkage; and (c) were not asked for consent to data linkage as they were not present or were interviewed by proxy in the FRS. The overall consent rate was 55%. There were only small differences between those who were asked for consent to data linkage and the overall PSE-UK working-age sample. Where there were differences, they were symmetrically mirrored among those who were not asked for consent. For example, young adults aged 18–24 years were under-represented among those who consented to data linkage (6% vs. 11%), while older adults aged 55–64 were over-represented (27% vs. 24%), and the reverse was found in the residual category. Similarly, female respondents were over-represented among those who were asked for consent; however, there was little difference between those who refused and those who gave consent for data linkage.

Individuals within households that were composed of either single adults, childless couples, couples where the youngest child was 10 years or younger, or lone parents were generally over-represented among those who consented to data linkage, while the same groups were under-represented among those who were not asked to provide consent. Individuals from a white ethnic background or who were residing in Scotland at the time of the survey were over-represented among those who consented to data linkage compared to the overall sample. These groups were under-represented among those who either refused or were not asked for consent. These findings broadly mirror those reported in an evaluation study of administrative data linkage of earlier 2009/2010 FRS data by McKay (2012). The author found a relatively low level of consent bias and remarkable similarities between consenters and non-consenters. The only observed differences were by household size and ethnic background for those who were interviewed by proxy in the FRS. In addition, the study found few differences in income levels between respondents who provided consent and the rest of the sample.

The remaining part of Table 1 shows that both over-representation among those who were asked to provide consent and under-representation among those who were not asked for consent are closely associated with the highest level of education, employment status and occupational category. For example, individuals with a higher educational level (i.e. degree and equivalent), along with those who were either employed or self-employed, or belonging to managerial or professional occupations were over-represented among those who consented to data linkage, and to a greater extent among those who refused to provide consent. The reverse was observed for the group who were not asked for consent. Within this group, individuals with an unknown educational level were substantially over-represented (42% vs. 22% in the overall sample), while those who were either permanently sick/disabled or with a longstanding illness/disability were substantially under-represented (2% vs. 5% of the overall sample in the former case; 12% vs. 26% in the latter case). No differences between the groups who either consented or were not asked for consent to data linkage were detected for unemployed individuals compared to the overall working-age sample.

Looking at the group which provided consent for data linkage, Table 2 shows the proportions where linkage was achieved to the different administrative data sources in the year of the PSE-UK survey or any of the preceding 4 years. More than a third (36%) of the final sample included any JSA or IS benefits records and more than two-thirds (70%) had information on earnings or employment from tax records. It was reassuring that approximately

Table 1 Composition of the PSE-UK working-age sample by those who: (a) consented to data linkage; (b) refused consent for data linkage; and (c) were not asked for consent to data linkage; unweighted data

Variable	PSE-UK working-age sample (n=5439)	(a) Consented to data linkage (n=2985)	(b) Refused consent for data linkage (n=1172)	(c) Not asked for consent (n=1282)
All	100	54.9	21.5	23.6
Age group (years)				
18–24	10.9	5.8	4.1	29.1
25–34	17.0	17.3	17.1	16.4
35–44	23.2	24.1	24.8	19.5
45–54	25.3	26.3	29.6	19.2
55–64	23.6	26.5	24.4	15.8
Gender				
Female	54.3	58.6	58.0	40.9
Male	45.7	41.4	42.0	59.1
Household composition				
Single	12.0	17.0	11.7	0.6
Couple	28.4	30.1	28.2	24.7
3+adults	12.4	8.7	10.7	22.6
Couple-youngest child 0–4 years	9.1	11.3	10.7	2.6
Couple-youngest child 5–10 years	7.8	9.2	10.0	2.6
Couple-youngest child 11+years	30.3	23.7	28.7	46.9
Lone parent household				
Yes	6.0	8.5	5.2	1.2
No	94.0	91.5	94.8	98.8
Ethnic group				
White	81.0	84.1	79.2	75.7
Not white	19.0	15.9	20.8	24.3
Country				
England and Wales	72.0	70.1	73.8	74.8
Scotland	28.0	29.9	26.2	25.2
Highest level of education				
Degree and equivalent	32.2	33.2	40.2	22.4
A Level and equivalent	9.8	9.8	10.7	9.1
O Level and CSE	25.4	29.0	24.6	17.8
Other/none	10.9	11.9	10.9	8.5
Not known	21.7	16.1	13.6	42.2
Employment status				
Employed	51.4	55.4	56.1	37.7
Self-employed	5.0	5.3	6.6	3.0
Unemployed	5.7	6.1	4.3	6.0
Permanently sick/disabled	5.3	7.3	4.4	1.5
Home carer	7.3	8.1	9.7	2.9
Other inactive	10.8	11.6	10.4	9.5
Not known	14.5	6.2	8.5	39.4

Table 1 (continued)

Variable	PSE-UK working-age sample (n=5439)	(a) Consented to data linkage (n=2985)	(b) Refused consent for data linkage (n=1172)	(c) Not asked for consent (n=1282)
Occupation (NS-SeC)				
Managerial and professional occupations	21.5	22.6	27.0	13.9
Intermediate occupations	10.3	11.1	10.5	8.0
Small employers and own account workers	3.7	4.0	4.1	2.9
Lower supervisory and technical occupations	3.2	3.3	3.9	2.4
Semi-routine and routine occupations	17.2	19.3	16.5	13.0
Not classified	44.1	39.7	38.0	59.8
Longstanding illness or disability				
Yes	25.9	31.9	25.5	12.2
No/not known	74.1	68.1	74.5	87.8

Table 2 Composition of the consented PSE-UK working-age sample, by linkage to administrative data sources

	n	%[a]
All consented	2985	100
Any benefits records[a]	1084	36.3
Any tax records[b]	2101	70.4
Either benefits and/or tax records	2221	74.4
Neither benefits nor tax records	764	25.6
Both benefits and tax records	964	32.3
Only benefits records	120	4.0
Only tax records	1901	63.7

[a]Benefits records include JSA and IS

[b]Tax records include employment spells from P45/P46 forms and annual earnings from P14/P60 forms

three quarters of the sample (74%) presented either benefits and/or tax records, while these were jointly absent for about one quarter of the sample (26%). However, it is worthwhile noting that the relatively high proportion of the final sample in receipt of any benefits records in the consented PSE-UK working-age sample may reflect the fact that the PSE-UK survey oversampled respondents from poorer households (Gordon 2011; pp. 10–11).

2.3 Validation Approach

Our approach to validating unemployment information is based on the comparison of information derived from survey data with corresponding spell information reconstructed from administrative sources. Cases with significant mismatch or misclassification between the two

sources were further investigated by comparing their demographic and employment profiles. Often studies comparing administrative and survey data take the administrative information as the 'gold standard', as it is collected for non-statistical purposes (e.g. Bruckmeier et al. 2018; Lynn et al. 2012; McKay 2012; Wichert and Wilke 2012). These studies tend to rest on the assumption that the information derived from administrative data is free from measurement error and is of higher quality compared to that collected from surveys. By contrast, our validation exercise does not entail any a priori assumptions on the level of accuracy of one data source over the other, as it is intended to illustrate the degree of concordance between the different sources and characteristics that may be associated with misclassification.

2.4 Modelling Approach and Variable Selection

In this study, we look at poverty status as our social disadvantage category, and we investigate how well longitudinal measures of labour market participation, derived from administrative and survey data, predict current poverty risks. As an outcome we use a poverty measure derived from PSE-UK which combines information on low income and deprivation: following a modelling exercise using the General Linear Model (i.e. ANOVA and logistic regression models), adults were identified as 'poor' if their equivalised household income after housing costs is below £295 per week and they lack three or more items on a deprivation scale identified using a consensual approach (Bailey and Bramley 2018; Gordon 2017).

The poverty outcome was missing in 11 cases reducing our analytical sample to 2974 respondents.

We modelled the binary poverty outcome with a logistic regression model. Denote by y_i the outcome variable of a respondent i ($i = 1, \ldots, n$), coded as

$$y_i = \begin{cases} 0 & not\, poor, \\ 1 & poor. \end{cases}$$

A logistic regression model for the log-odds of being poor versus not being poor may be written as:

$$\log\left[\frac{\pi_i}{1 - \pi_i}\right] = \mathbf{x}'_i \boldsymbol{\beta} \tag{1}$$

where π_i is the probability of being 'poor', $\mathbf{x}'_i$ is a vector of observed characteristics of the respondents and their households and $\boldsymbol{\beta}$ is a vector of regression coefficients.

From the PSE-UK survey, we have self-reported employment status at the survey date and hence unemployment, as well as unemployment over the last 5 years from a recall question capturing duration in months. From administrative data, we have receipt of JSA and/or IS at the survey date as well as the sum of unemployment spells over the 5 calendar years prior to PSE-UK survey date. A similar longitudinal measure for earnings from employment over the 5 years prior to the survey was reconstructed from HMRC tax records from P14/P60 forms. Gross annual earnings for a given tax year were calculated by summing earnings across all employments. As employment spell dates (imported from P45/46 forms) had a high rate of missing observations, earnings were allocated to calendar years assuming a uniform distribution over each tax year. Control variables were mostly derived from PSE-UK and included: age, gender, household composition, whether lone parent household, ethnic group, country of residence, highest level of education and whether respondent had a longstanding illness or disability.

Table 3 Comparison of number of months unemployed over the last 5 years from survey and administrative benefits data, PSE-UK 2012/WPLS linked (unweighted) data

| Months unemployed last 5 years (survey) | Months in receipt of Jobseeker's Allowance (JSA) and/or Income Support (IS) last 5 years (administrative benefits data; n=2985) | | | | 0 months in receipt of JSA and/or IS (n=2348) | |
| | | | | | Recorded in administrative benefits data | |
	0 months (n=2348) %	<6 months (n=178) %	6–11 months (n=111) %	12+months (n=348) %	Yes (n=447) %	No (n=1901) %
0 months	**89.3**	47.2	36.0	47.4	85.9	90.1
<6 months	3.1	**21.9**	8.1	3.2	3.8	2.9
6–11 months	1.7	11.8	**21.6**	2.6	0.7	1.9
12+months	5.9	19.1	34.3	**46.8**	9.6	5.1
Total	100	100	100	100	100	100

Bolded values are diagonal cells indicating the proportion of cases for corresponding categories for the variables derived from survey and administrative benefits data

For the estimation of the multivariate models we used the 'svyset' and 'subpop' commands in Stata/SE 14.2 (StataCorp LP, College Station, Texas) to account for the complex design of the PSE-UK survey. All models were estimated with robust standard errors to reflect the survey design. Wald test statistics for the key variables of interest were reported to inform whether the parameters associated with a specific explanatory variable were significantly different from zero and the variable should be therefore included in the model. The summary statistics of the variables included in the models are presented in Table 7 in the "Appendix".

3 Results

In this section, we begin by addressing the first of the preliminary research questions by comparing responses from the survey with the equivalent measures derived from administrative data. We then present the descriptive results from the pre-modelling stage of statistical analyses before discussing the final selected models, which address the remaining research questions outlined in Sect. 1.

3.1 Internal Validation

Overall, our analyses suggest that the survey and administrative sources are very comparable, and where they differ, it has more to do with differences in definitions or interpretation than error due to mismatching of records to surveys. McKay (2012; p. 31) reaches similar conclusions and observes that combining administrative data with survey data provides more accurate results than when using standard survey data.

We first compare the two measures of number of months unemployed over the last 5 years: from the survey recall question in PSE-UK and from the equivalent measure constructed from administrative data (Table 3). The rate of correspondence between the two

Table 4 Composition of the sub-sample of those reporting zero months unemployed in the last 5 years in PSE-UK survey, by months of receipt of Jobseeker's Allowance and/or Income Support from administrative benefits data; PSE-UK 2012/WPLS linked (unweighted) data

| Variable | Months in receipt of Jobseeker's Allowance (JSA) and/or Income Support (IS) last 5 years (administrative benefits data; n=2386) | | | |
	0 months (n=2097) %	<6 months (n=84) %	6–11 months (n=40) %	12+months (n=165) %
Employment status				
Employed	63.2	48.8	37.5	11.5
Self-employed	5.8	0.0	0.0	1.2
Unemployed	1.4	4.9	2.5	5.5
Sick/disabled	4.0	7.1	7.5	44.2
Home carer	6.4	7.1	15.0	18.2
Other inactive	12.3	10.7	17.5	12.1
Not known	6.9	21.4	20.0	8.5
Total	100	100	100	100

measures is highest (89%) for those who did not experience any episode of unemployment in that time. The second highest rate is observed for those who were unemployed for 12 months or longer but here the correspondence rate is just 47%, and it is lower still for the intermediate categories. The main reason for the reduced rates of matching is the high proportions with recorded episodes of unemployment but who report no unemployment when asked in the survey (between 36% and 47% in each case).

These cases are further scrutinised in Table 4, which shows their composition by employment status, as reported in the PSE-UK survey. The main reason for the discrepancy between the measures appears to be that many of those not in employment do not regard themselves as being unemployed but rather as economically inactive: at the time of the survey, they report themselves as sick/disabled (44%) or as home carers (18%) when they have been in receipt of JSA and/or IS for 12 months or longer. If we were to break down these categories by gender, we would see that women are more likely to be home carers whereas men are more likely to be sick or disabled (not shown in Table 4). This is in line with the results of a study by Dex and McCulloch (1998) based on a comparison of two British longitudinal survey data sources. The authors find that when analysing retrospective unemployment history information, women appear to be affected to a larger extent by definitional issues, as they often have caring responsibilities and tend to perceive themselves as inactive rather than unemployed, compared to men (p. 505). Returning to Table 4, very few of those in receipt of JSA and/or IS for 12 months or longer, reported themselves as employed (12%) or self-employed (1%). This provides reassurance that the administrative data measure is still picking up people not in employment, and hence at greater risk of unemployment, even if they do not regard themselves as having been 'unemployed'. This also highlights some of the difficulties in identifying (non-)employment status from welfare benefits data.

Returning to Table 3, the last two columns focus on those with no recorded JSA/IS claims in the previous 5 years (i.e. those covered by the first column of this table). We are concerned here with whether this group may have low recorded unemployment
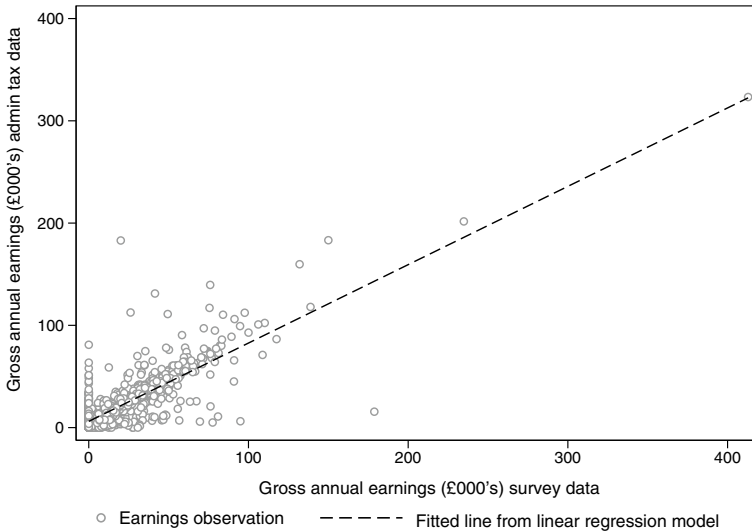
Fig. 2 Comparison of gross annual earnings from FRS 2010/2011 survey and HMRC administrative data for the tax year 2010/2011[a] (r = 0.84, n = 1815). *Note* [a] Earnings from administrative tax records are gross of all statutory and non-statutory deductions except for occupational pension contributions made by employees before tax

because of a failure to match data rather than a genuine absence in spells on unemployment benefits. We therefore separate those who have some recorded claims in our records, albeit none in the 5 years before the survey, from those with no claims records at all. The group with no records is much larger (1901 compared with 447) but, in both cases, the overwhelming majority are people who also self-report no unemployment. This provides strong reassurance that the absence of a record does indicate a lack of unemployment rather than a failure to match records.

Similar comparisons can be made of gross annual earnings recorded in the HMRC data and the gross earnings reported in the FRS survey. The latter are collected as weekly or monthly figures and converted to an annual measure. Earnings were not recorded in the PSE-UK as it did not duplicate information already collected in the earlier survey (see Sect. 2.1).

Despite the fact that the recorded measure in the HMRC data is annual and the reported one in the FRS survey is annualised, the two measures are strongly correlated (r = 0.84) although a number of outliers can be observed (Fig. 2). In nearly 4 out of 10 cases, where respondents reported some earnings, there were no administrative data. In addition, we observe that for respondents reporting zero earnings in the FRS, there is a range of values for recorded earnings up to £100,000 (with missing values accounting for more than 5 out of 10 cases when FRS reported earnings are equal to zero), which suggest possible imputation or editing performed on FRS values. In line with findings reported by McKay (2012), these range of values together with observations located above the fitted line in proximity of the y axis denote under-reporting of FRS compared to administrative tax data.

Table 5 PSE-UK poverty rates by key variables from survey and administrative data, PSE-UK 2012/WPLS linked (weighted) data

Variable	PSE-UK poverty rates			
	Survey (%)	Weighted count All (n)	Administrative data (%)	Weighted count All (n)
Unemployed[a] at survey date				
No	26.0	3612.0	25.6	3586.0
Yes	74.3	260.1	74.3	286.4
F-test statistic		97.2		148.3
(df1; df2)		(1; 1127)		(1; 1127)
Months unemployed[a] last 5 years				
0 months	22.1	3025.0	22.0	3111.0
<6 months	31.1	224.8	41.5	237.6
6–11 months	50.6	152.6	47.7	126.4
12 + months	67.3	470.0	72.2	397.7
F-test statistic		50.1		63.2
(df1; df2)		(2.9; 3314.6)		(2.9; 3257.2)
Earnings last 5 years				
0 earnings	—	—	35.3	1486.0
Lower tertile	—	—	45.6	744.8
Middle tertile	—	—	26.1	796.9
Higher tertile	—	—	7.1	845.2
F-test statistic		—		40.7
(df1; df2)		—		(2.9; 3311.9)

Note: Percentages and observations count are corrected for survey design; *F*-test statistic and degrees of freedom for numerator and denominator (df1; df2, respectively) are reported for a Pearson χ^2 test of independence, corrected for survey design with the Rao and Scott (1984) second-order correction, *p* value < 0.001; the overall estimated proportion of poor individuals in the analytic sub-sample of reference is 29.2% (n = 1131)

[a]'Unemployed' to be read as 'self-reported unemployed' in the case of survey and 'in receipt of JSA and/or IS' in the case of administrative data

3.2 Descriptive Results

We now move to compare how well survey and administrative measures of employment status or history predict current poverty risks. Table 5 shows poverty rates by current employment status and recent employment history as measured by the two sources. It uses the PSE-UK poverty measure, which combines low income and deprivation, as described in Sect. 2.4. For current employment status, the survey measure of (self-defined) unemployment is compared with the administrative data measure of receipt of JSA and/or IS. For employment history, self-reported months unemployed in the previous 5 years is compared with months of JSA and/or IS receipt during the same period. The table also shows poverty rates by earnings over the last 5 years which was available only from the administrative data. Where individuals have some earnings, we divide them into three groups (tertiles) but we also show those with no recorded earnings separately.

As expected, individuals unemployed at survey date and with more unemployment during the last 5 years had higher poverty rates, regardless of whether we use survey or

administrative measures. Differences between poor and not poor individuals were detected by an F-test statistic for a Pearson χ^2 test of independence, corrected for survey design. Poor individuals differed significantly from those not poor across all variables of interest. With the earnings measure for the last 5 years, there is a clear gradient with poverty risks across the tertiles but the group with no recorded earnings has slightly lower poverty risks than those with the lowest tertile. That suggests that this group includes a mix of people unable to work or to find work, but also others who might be self-employed or choose not to work.

3.3 Results of Logistic Regression Modelling

Table 6 presents the results of a set of logistic models. We started by estimating a baseline model (Model 1) which included all the control variables. Models 2a–2d allow us to address the second research question, on how well our administrative data measures predict poverty risks compared with the equivalent survey measures. Model 3 addresses the third research question on the extent to which longitudinal measures of unemployment add to our understanding of poverty risks, compared with current measures. Lastly, Model 4 let us examine the extent to which longitudinal measures of earnings add further to our understanding of poverty risks, even after controls for unemployment histories. We use the Taylor series linearization, a sandwich-type variance estimating method, to obtain more accurate or robust standard errors accounting for the fact that individuals are clustered in households (for details see Heeringa et al. (2017), pp. 70–75). To ease interpretation, the estimated average marginal effects of the variables of interest on the probability of being in poverty are plotted in Figs. 3, 4 and 5, indicating the effect of a discrete change of a variable of interest on the outcome, calculated as the average on the predicted values generated by the main regression models presented in Table 6. In the "Appendix" we also report the results obtained from logistic regression models not accounting for survey design (Table 8). The results are broadly in line with those presented in Table 6. Although model-fit statistics slightly favour models including longitudinal unemployment information derived from survey data, when considering models including longitudinal measures reconstructed from administrative data sources, the results in both tables are pointing in the same direction.

In the baseline model (Model 1), relationships with control variables were very much as expected. The odds of being in poverty declined with age, reflecting progression in the labour market. Households with children have higher poverty risks than childless couples, with risks increasing where children are younger and more constraining on employment. They were lower for women than men but this is after controlling for household composition and lone parent status; poverty risks are shown to be much higher for lone parents, the vast majority of whom are women. Once other factors have been controlled for, poverty risks did not vary with ethnicity but residing in Scotland (vs. England and Wales) was associated with lower risks of poverty (see Bailey (2014) for a discussion). Higher qualifications and higher status occupational classes were associated with lower poverty, reflecting higher earnings. A long-standing illness or disability was associated with higher poverty risks. These results were largely stable across subsequent models.

Table 6 Odd ratios and standard errors of logistic regression models of PSE-UK poverty status accounting for survey design, PSE-UK 2012/WPLS linked data

Variable	Model1		Model 2a		Model 2b		Model 2c		Model 2d		Model 3		Model 4	
	OR	SE	OR	SE	OR	SE	OR	SE	OR	SE	OR	SE	OR	SE
Age group (reference: 55-64 years)														
18-24	5.66***	1.71	4.63***	1.42	5.09***	1.56	3.98***	1.19	4.74***	1.49	4.68***	1.47	4.03***	1.25
25-34	3.04***	0.71	2.70***	0.63	2.84***	0.66	2.37***	0.58	2.60***	0.61	2.61***	0.61	2.22*	0.53
35-44	3.06***	0.73	2.74***	0.65	2.83***	0.67	2.64***	0.64	2.68***	0.64	2.67***	0.64	2.47***	0.59
45-54	1.82**	0.35	1.63*	0.31	1.64*	0.32	1.58*	0.30	1.60*	0.30	1.57*	0.30	1.56*	0.29
Gender (reference: male)														
Female	0.72**	0.10	0.74*	0.10	0.75*	0.11	0.76†	0.11	0.76†	0.11	0.77†	0.11	0.69*	0.10
Household composition (reference: couple with no children)														
Single	2.73***	0.54	2.61***	0.53	2.56***	0.52	2.52***	0.53	2.44***	0.50	2.42**	0.50	2.35**	0.48
3+ adults	0.65	0.23	0.72	0.26	0.70	0.25	0.73	0.26	0.70	0.27	0.71	0.27	0.72	0.27
Couple -youngest child 0-4 years	3.93***	0.96	4.15***	1.00	3.93***	0.95	4.17***	1.04	3.85***	0.94	3.85***	0.94	4.07***	1.01
Couple -youngest child 5-10 years	3.41***	0.94	3.51***	0.97	3.31***	0.91	3.48***	0.99	3.37***	0.92	3.34***	0.91	3.63***	1.01
Couple -youngest child 11+ years	2.05**	0.44	2.14***	0.46	2.04**	0.43	2.29***	0.50	2.15***	0.46	2.13***	0.46	2.28***	0.48
Lone parent household (reference: no)														
Yes	2.61**	0.75	2.56**	0.74	2.42**	0.70	2.44**	0.70	2.18**	0.64	2.19**	0.65	2.12*	0.63
Ethnic group (reference: white)														
Not white	1.35	0.25	1.32	0.23	1.39†	0.26	1.33	0.24	1.43†	0.27	1.43†	0.27	1.34	0.25
Country (reference: England and Wales)														
Scotland	0.72*	0.12	0.72*	0.12	0.72*	0.12	0.71*	0.12	0.70*	0.12	0.70*	0.12	0.71*	0.12
Highest level of education (reference: none/other)														
Degree and equivalent	0.33***	0.08	0.34***	0.08	0.34***	0.08	0.33***	0.08	0.34***	0.08	0.35***	0.08	0.36***	0.09
A level and equivalent	0.54†	0.18	0.61	0.20	0.59	0.20	0.59	0.20	0.62	0.21	0.62	0.21	0.60	0.21
O level and CSE	1.02	0.23	1.06	0.25	1.04	0.24	1.05	0.25	1.06	0.25	1.06	0.25	1.07	0.25
Not known	1.94**	0.47	2.09**	0.52	1.86*	0.46	2.10**	0.54	1.85*	0.47	1.84*	0.47	1.76*	0.45
Occupation (NS-Sec) (reference: semi-routine and routine occupations)														
Managerial and professional occupations	0.23***	0.06	0.24***	0.06	0.24***	0.06	0.29***	0.08	0.26***	0.07	0.26***	0.07	0.32***	0.08
Intermediate occupations	0.57**	0.14	0.56*	0.14	0.57*	0.14	0.65†	0.16	0.59†	0.15	0.59†	0.15	0.65†	0.16
Small employers and own account workers	0.46**	0.16	0.47*	0.16	0.47*	0.17	0.56†	0.19	0.51†	0.18	0.51†	0.18	0.46*	0.17
Lower supervisory and technical occupations	0.62	0.25	0.63	0.25	0.62	0.23	0.78	0.30	0.67	0.25	0.67	0.25	0.78	0.28
Not classified	2.43***	0.44	1.90***	0.35	2.09***	0.38	2.33***	0.43	2.12***	0.38	2.06***	0.37	1.99***	0.37
Long-standing illness or disability (reference: no/not known)														
Yes	1.87***	0.28	1.95***	0.30	1.87***	0.28	1.87	0.28	1.81***	0.27	1.83***	0.27	1.79***	0.26

Table 6 (continued)

Variable	Model1		Model 2a		Model 2b		Model 2c		Model 2d		Model 3		Model 4	
	OR	SE	OR	SE	OR	SE	OR	SE	OR	SE	OR	SE	OR	SE
Unemployed at survey date (survey) (reference: no)														
Yes			3.03***	0.82										
In receipt of JSA and/or IS (administrative data) (reference: no)														
Yes					2.30***	0.54					1.38	0.44		
Months unemployed last 5 years (survey) (reference: 0 months)														
<6 months							1.73†	0.53						
6–11 months							4.34***	1.31						
12+ months							3.15***	0.60						
Months of JSA and/or IS receipt last 5 years (administrative data) (reference: 0 months)														
<6 months									1.66†	0.44	1.60†	0.43	1.72*	0.46
6–11 months									2.43**	0.71	2.25**	0.69	2.40**	0.73
12+ months									2.30***	0.51	1.93†	0.57	2.20***	0.48
Earnings last 5 years (administrative data) (reference: higher tertile)														
0 earnings													2.88***	0.69
Lower tertile													2.44**	0.68
Middle tertile													2.35***	0.63
Intercept	0.11***	0.03	0.11***	0.04	0.11***	0.04	0.09***	0.03	0.10***	0.03	0.10***	0.03	0.05***	0.02
Observations	2,897		2,897		2,897		2,897		2,897		2,897		2,897	
Clusters	1,138		1,138		1,138		1,138		1,138		1,138		1,138	
Strata	11		11		11		11		11		11		11	
Overall F-test statistic	15.59***		15.00***		14.99***		14.60***		14.30***		13.83***		13.20***	
(df1; df2)	(23; 1105)		(24; 1104)		(24; 1104)		(26; 1102)		(26; 1102)		(27; 1101)		(29; 1099)	
F-test statistic of additional covariates[b]			16.97***		12.57***		17.34***		7.08***		5.91***		7.27***	
(df1; df2)			(1; 1127)		(1; 1127)		(3; 1125)		(3; 1125)		(4; 1124)		(6; 1122)	

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; † $p < 0.10$

Taylor Series Linearized standard errors are provided

[a]The F-test statistics refers to an adjusted Wald test for nested models accounting for complex survey design including additional covariates compared to a baseline model (Model 1). F-test statistics for additional comparisons are reported as follows: Model 3 versus Model 2b: 3.53* (df1 = 3; df2 = 1125); Model 3 versus Model 2d: 1.07 (df1 = 1; df2 = 1127); Model 4 vs. Model 2d: 6.54*** (df1 = 3; df2 = 1125). When accounting for complex survey design, standard model-fit statistics such as Likelihood ratio test, AIC and BIC cannot be computed
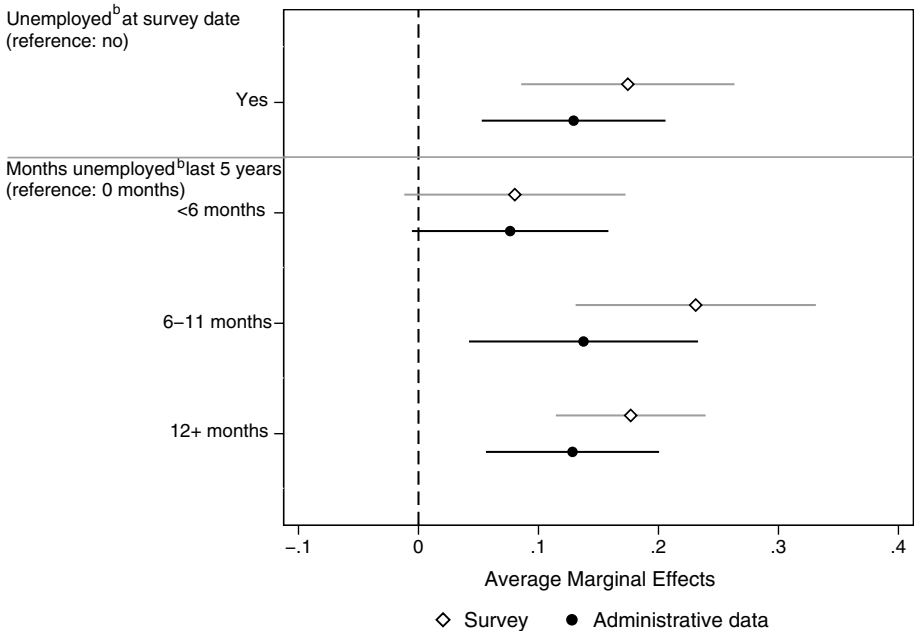
Fig. 3 Average marginal effects[a] with 95% confidence intervals of the probability of PSE-UK poverty by cross-sectional and longitudinal measures of unemployment from survey and administrative data; PSE-UK 2012/WPLS linked data. *Notes* [a] Average marginal effects are obtained from Models 2a–2d accounting for survey design (Table 6), additionally controlling for: age groups, gender, household composition, whether lone parent household, country, highest level of education, occupation and disability/long-term illness. [b] 'Unemployed' to be read as 'self-reported unemployed' in the case of survey and 'in receipt of JSA and/or IS' in the case of administrative data

Presented in Models 2a and 2b are the associations of current unemployment with poverty risks using survey and administrative data measures, respectively. The odds of being in poverty are much higher for those unemployed at the survey and by similar amounts. As Fig. 3 shows, there is considerable overlap in the estimates. Turning to the longitudinal measures of unemployment, there is again consistency, with poverty risks, that are higher for those with more unemployment in both cases. The differences appear more marked with the self-reported measure but confidence intervals are also wider (with significance levels varying between $p < 0.10$ and $p < 0.001$) and there is considerable overlap in the estimated associations with poverty (Fig. 3). Furthermore, the greatest poverty risks are observed for individuals unemployed for 6–11 months. With longer unemployment spells, poverty risks drop slightly suggesting that this group contains some people who do not need to work, perhaps because they have a partner with reasonable earnings. In all four models, the inclusion of the additional covariates significantly improves model fit compared to the baseline model. Based on the values of the *F*-statistic, corresponding to a Wald test adjusted for survey design, we can reject the null hypothesis that each additional covariate is equal to zero at $p < 0.001$.

Model 3 (and Fig. 4) include both longitudinal and current measures of unemployment from administrative benefits data. When we include the longitudinal measure, the measure of current benefits receipt has a much weaker relationship and, indeed, is no longer
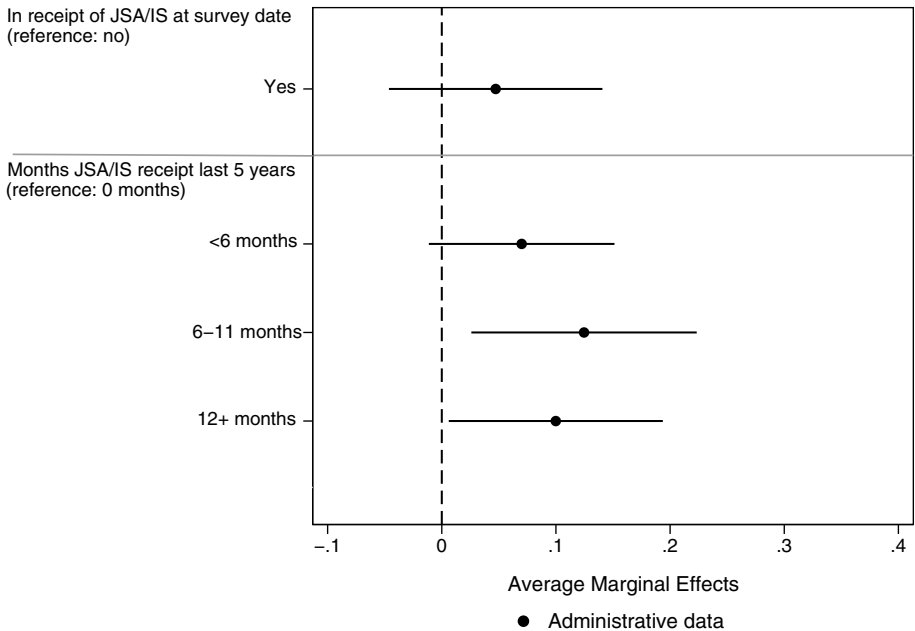
Fig. 4 Average marginal effects[a] with 95% confidence intervals of the probability of PSE-UK poverty by cross-sectional and longitudinal measures of receipt of Jobseeker's Allowance (JSA) and/or Income Support (IS); PSE-UK 2012/WPLS linked data. *Note*: [a] Average marginal effects are obtained from Model 3 accounting for survey design (Table 6), additionally controlling for: age groups, gender, household composition, whether lone parent household, country, highest level of education, occupation and disability/long-term illness

significant. The relationship with longitudinal unemployment has a similar form, with the greatest poverty risks for those unemployed for 6–11 months. If we compare the overall fit of this model with the one with current unemployment alone (Model 2b), we see that the inclusion of the longitudinal covariate leads to a significant improvement (the *F*-statistic for the adjusted Wald test is 3.53; 3 and 1125 degrees of freedom; $p < 0.05$; see Table 6). Alternatively, if we compare the overall fit of the same model with the one including the longitudinal measure of unemployment on its own (Model 2d), we see that by adding current unemployment we do not obtain any significant improvement (the *F*-statistic for the adjusted Wald test is 1.07; 1 and 1127 degrees of freedom; $p > 0.10$; see Table 6). In other words, current unemployment does not add to our understanding of unemployment based on the longitudinal measure.

Lastly, we turn to Model 4 (and Fig. 5) where we begin to explore the advantages of bringing in administrative measures from other domains, in this case earnings from employment. The addition of the longitudinal measure of earnings from employment significantly improved the model fit, not only when compared to the baseline model (Model 1) but also when compared to the model with the longitudinal unemployment measure on its own (Model 2d; the *F*-statistic for the adjusted Wald test is 6.54; 3 and 1125 degrees of freedom; $p < 0.001$; see Table 6). Once we control for unemployment history, those with no earnings now have the higher poverty risks, greater than those with earnings in the lowest tertile. Overall, the influence of earnings history appears as great as the influence of unemployment history.
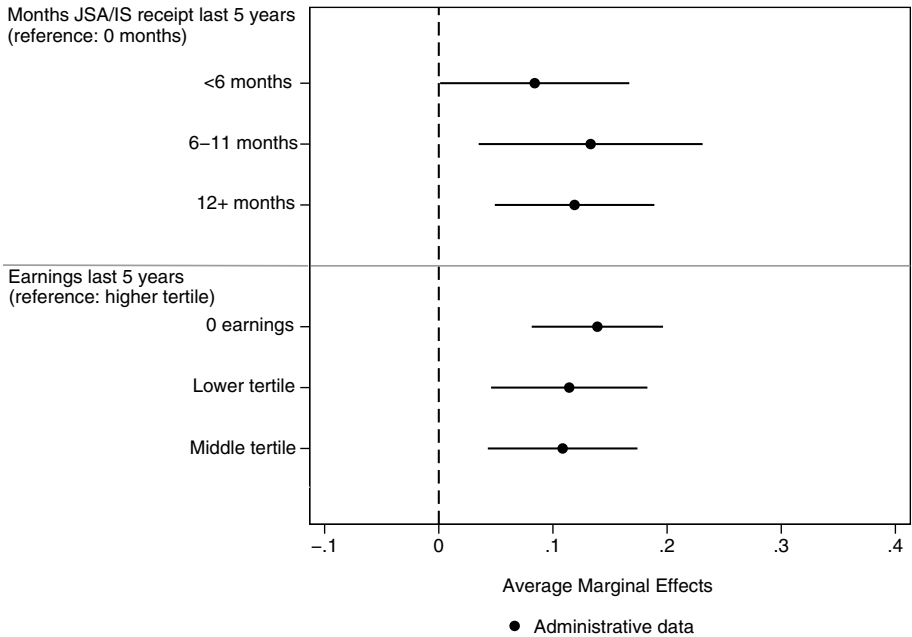
Fig. 5 Average marginal effects[a] with 95% confidence intervals of the probability of PSE poverty by longitudinal administrative measures of receipt of Jobseeker's Allowance (JSA) and/or Income Support (IS) and earnings; PSE-UK 2012/WPLS linked data. *Note*: [a] Average marginal effects are obtained from Model 4 accounting for survey design (Table 6), additionally controlling for: age groups, gender, household composition, whether lone parent household, country, highest level of education, occupation and disability/long-term illness

4 Summary and Discussion

Social indicators such as Free School Meals eligibility and Indices of Multiple Deprivation, which have been developed in policy and for operational purposes, are frequently taken up by researchers wishing to investigate the impacts of social disadvantage on outcomes such as educational attainment and inequalities both within and across neighbourhoods. Yet the design of these indicators is heavily compromised as they are, necessarily, a tradeoff between accuracy and simplicity or ease of application. FSM eligibility, for example, is based on snapshot indicators from one domain (welfare benefits receipt) and these indicators are also at the core of composite indices such as IMDs which combine indicators from multiple domains. This paper explores the potential for researchers to devise much more accurate measures, taking advantage of the improving access to administrative data. Our aim is not to develop an ideal measure: the *best* measure for any given research will depend on the question being addressed. Rather the aim is to demonstrate the potential of this approach.

To this end, we used individual-level administrative data from welfare and tax systems on benefits receipt and earnings from employment over a 5-year period, which were linked to survey data from a large national survey of poverty and social exclusion. We compared self-reported information from survey data with corresponding information recorded in administrative data to validate both current status and longitudinal information on unemployment and earnings. By using logistic regression models allowing for survey design, we compared indicators from survey and administrative data and investigate whether longitudinal

administrative measures capturing labour market histories are stronger predictors of current poverty risks than cross-sectional measures. In particular, we explored the advantages of moving from cross-sectional to longitudinal measures from a single source, and the advantages of moving from a single domain (welfare benefits) to two domains (adding earnings).

Using our linked survey and administrative dataset, our descriptive analyses show that there is good correspondence between measures of labour market history and earnings across these two sources. Divergences between the measures do not appear to be the result of problems with the linkage process as they are not random but systematic. Furthermore, these systematic differences can be reasonably explained by differences in the way the measures were constructed (annual recorded earnings vs. weekly/monthly self-reported earnings) or by different interpretations or understandings of concepts such as employment status. In the latter case, the benefits system takes a very specific definition of 'unemployment' to judge eligibility for JSA. While individual claimants must meet the relevant criteria to claim the benefit, this may not be how they see themselves. As a result, the survey records their status differently, in the 'economically inactive' category rather than 'unemployed'. They claim an unemployment benefit, JSA, because of the restricted alternatives, but that is not how they would define themselves so the benefits system *misrepresents* their labour market status.

Despite the very limited range of welfare benefits data, our regression analyses show that the administrative data measures performed as well as survey measures in predicting current poverty risks. This is corroborated by the examination of model-fit statistics for models including both measures of current status and longitudinal measures capturing previous unemployment. In the latter case, it was striking that both survey and administrative data measures showed the same unexpected, non-linear relationship with poverty risks. We also showed how increasing the complexity of administrative data measures improves prediction of poverty risks, without loss of information. As demonstrated by measures of improvement of model fit, there is a gain both in the move from cross-sectional to longitudinal measures within one domain, and a further gain when bringing in a second domain (earnings).

There are obvious limitations in the measures we have devised here. The clearest is that the coverage of welfare benefits data is limited to JSA/IS due to practical constraints. Extending the coverage to include other out-of-work benefits (Employment and Support Allowance, most obviously) or in-work tax credits would clearly improve the accuracy of our administrative data measures. The point is that, even with our limited data, we can show gains from moving from cross-sectional to longitudinal measures. Another issue is that analyses reported here were conducted at individual level, whereas poverty risks are substantially influenced by household circumstances (e.g. Maître et al. 2003). For the understanding of poverty, further work could explore what new light is shed through the information provided by administrative data on labour market histories. Again, the point of the current analysis was to demonstrate the value of the method, rather than address a substantive question. An additional issue, partly connected to the previous ones, is the fact that our outcome is a poverty measure which is derived from survey data, by combining information on low income and deprivation. It is likely that the low income component of this measure suffers from some measurement error. As mentioned above, we only have limited information on both earnings and benefit payments from administrative data sources which currently do not allow us to reconstruct a low income measure from household-level information. This is an area that certainly needs to be investigated. Further work could assess how low income measures derived from survey data compare to equivalent measures derived from administrative data.

There are many directions for further research. In general, the linkage of administrative data to the annual FRS data has not been sufficiently exploited, perhaps due to the

difficulties in accessing the data. Information from HMRC tax data could be used to study in more detail the relationships between employment spells, earnings and individual poverty risks, and to bring a household dimension into these. It would also provide the opportunity to contribute to the existing literature on the validation of employment and earnings survey measures (e.g. Francesconi et al. 2011; Lynn et al. 2012). From a methodological perspective, combining more recent survey collections with administrative data on a wider pool of welfare benefits would enable us to consider and address the consequences of misclassification and measurement errors arising from the linkage of administrative and survey data (Gilbert et al. 2017; Harron et al. 2017b).

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

See Tables 7 and 8.

Table 7 Summary statistics of the independent variables by PSE-UK poverty status, PSE-UK 2012/WPLS linked data

PSE-UK poverty status:	Not poor (n = 2742)		Poor (n = 1131)	
Variable	%	Weighted count n	%	Weighted count n
Age group (years)				
18–24	8.4	229.7	17.3	195.5
25–34	22.5	617.8	24.8	280.5
35–44	21.2	582.4	24.7	279.8
45–54	23.7	649.8	19.2	216.9
55–64	24.2	662.0	14.0	158.2
F-test statistic				7.8
(df1; df2)				(3.1; 3346.5)
Gender				
Female	52.1	1429.0	54.4	615.1
Male	47.9	1312.0	45.6	515.7

Table 7 (continued)

PSE-UK poverty status:	Not poor (n=2742)		Poor (n=1131)	
Variable	%	Weighted count n	%	Weighted count n
F-test statistic				0.7
(df1; df2)				(1.0;1070.0)
Household composition				
Single	16.1	441.1	22.0	248.5
Couple	37.0	1013.0	15.9	180.6
3 + adults	12.2	335.1	8.0	90.7
Couple-youngest child 0–4 years	9.1	250.2	20.4	230.3
Couple-youngest child 5–10 years	5.4	148.9	10.9	123.3
Couple-youngest child 11 + years	20.2	552.8	22.8	257.4
F-test statistic				12.3
(df1; df2)				(2.9;3083.0)
Lone parent household				
Yes	3.6	98.8	14.2	160.0
No	96.4	2643.0	85.8	970.8
F-test statistic				51.7
(df1; df2)				(1.0;1070.0)
Ethnic group				
White	91.3	2503.0	88.3	998.1
Not white	8.7	238.9	11.7	132.7
F-test statistic				4.3
(df1; df2)				(1.0;1070.0)
Country				
England and Wales	89.8	2462.0	92.0	1040.0
Scotland	10.2	279.5	8.0	90.1
F-test statistic				3.2
(df1; df2)				(1.0;1070.0)
Highest level of education				
Degree and equivalent	41.1	1128.0	11.9	135.0
A Level and equivalent	12.1	330.1	9.7	109.8
O Level and CSE	28.6	784.8	41.5	469.6
Other/none	9.8	268.2	13.3	149.8
Not known	8.4	230.8	23.6	266.6
F-test statistic				26.3
(df1; df2)				(2.5;2673.8)
Longstanding illness or disability				
Yes	25.6	701.8	41.2	466.2
No/not known	74.4	2040.0	58.8	664.6
F-test statistic				32.8
(df1; df2)				(1.0;1070.0)
Occupation (NS-SeC)				
Managerial and professional occupations	34.4	943.9	6.2	69.8
Intermediate occupations	15.9	434.9	8.2	92.9

Table 7 (continued)

PSE-UK poverty status:	Not poor (n=2742)		Poor (n=1131)	
Variable	%	Weighted count n	%	Weighted count n
Small employers and own account workers	4.7	128.0	2.7	31.0
Lower supervisory and technical occupations	4.1	113.0	2.7	30.9
Semi-routine and routine occupations	17.3	473.7	23.0	259.9
Not classified	23.6	648.0	57.2	646.2
F-test statistic				47.8
(df1; df2)				(4.8;5167.8)
Unemployed at survey data (survey)				
Yes	2.4	66.9	17.1	193.3
No	97.6	2675.0	82.9	937.5
F-test statistic				97.2
(df1; df2)				(1.0;1070.0)
In receipt of JSA and/or IS (administrative data)				
Yes	2.7	73.7	18.8	212.6
No	97.3	2668.0	81.2	918.2
F-test statistic				148.3
(df1; df2)				(1.0;1070.0)
Months unemployed last 5 years (survey)				
0 months	86.0	2358.0	59.0	667.2
<6 months	5.6	154.9	6.2	69.9
6–11 months	2.8	75.3	6.8	77.3
12+months	5.6	153.6	28.0	316.4
F-test statistic				50.1
(df1; df2)				(2.9;3146.8)
Months of JSA and/or IS last 5 years (administrative data)				
0 months	88.5	2426.0	60.6	684.7
<6 months	5.1	138.9	8.7	98.7
6–11 months	2.4	66.1	5.3	60.4
12+months	4.0	110.5	25.4	287.2
F-test statistic				63.2
(df1; df2)				(2.9;3092.3)
Earnings last 5 years (administrative data)				
0 earnings	35.1	961.7	46.3	523.9
Lower tertile	14.8	405.4	30.0	339.4
Middle tertile	21.5	589.1	18.4	207.8
Higher tertile	28.7	785.4	5.3	59.7
F-test statistic				40.7
(df1; df2)				(2.9;3144.4)

Note: Percentages and observations count are corrected for survey design; *F*-test statistic and degrees of freedom for numerator and denominator (df1; df2, respectively) are reported for a Pearson χ^2 test of independence, corrected for survey design with the Rao and Scott (1984) second-order correction, *p*-value <0.001

Table 8 Odd ratios and standard errors of logistic regression models of PSE-UK poverty status not accounting for survey design, PSE-UK 2012/WPLS linked data

Variable	Model1		Model 2a		Model 2b		Model 2c		Model 2d		Model 3		Model 4	
	OR	SE	OR	SE	OR	SE	OR	SE	OR	SE	OR	SE	OR	SE
Age group (reference: 55-64 years)														
18-24	3.64***	0.83	3.09***	0.72	3.24***	0.74	2.64***	0.63	2.93***	0.68	2.90***	0.67	2.45***	0.57
25-34	2.41***	0.42	2.08***	0.37	2.26***	0.40	1.91***	0.34	2.02***	0.36	2.02***	0.36	1.77*	0.32
35-44	2.29***	0.36	2.03***	0.33	2.13***	0.34	1.94***	0.32	1.99***	0.32	1.98***	0.32	1.85***	0.30
45-54	1.50**	0.22	1.31†	0.19	1.36*	0.20	1.32†	0.20	1.31†	0.19	1.30†	0.19	1.26	0.19
Gender (reference: male)														
Female	0.97	0.10	1.02	0.11	1.01	0.10	1.04	0.11	1.05	0.11	1.05	0.11	0.96	0.10
Household composition (reference: couple with no children)														
Single	2.53***	0.36	2.43***	0.35	2.34***	0.34	2.38***	0.35	2.19***	0.32	2.19***	0.32	2.18***	0.32
3+ adults	0.53**	0.12	0.56**	0.12	0.56**	0.12	0.60*	0.13	0.55**	0.12	0.56**	0.12	0.57*	0.13
Couple -youngest child 0-4 years	3.25***	0.60	3.45***	0.64	3.26***	0.60	3.51***	0.67	3.25***	0.61	3.25***	0.61	3.30***	0.62
Couple -youngest child 5-10 years	3.60***	0.71	3.73***	0.75	3.48***	0.70	3.94***	0.81	3.54***	0.71	3.53***	0.71	3.72***	0.76
Couple -youngest child 11+years	1.78***	0.26	1.89***	0.29	1.79***	0.27	2.07***	0.32	1.87***	0.28	1.86***	0.28	1.96**	0.30
Lone parent household (reference: no)														
Yes	3.12***	0.56	2.97***	0.54	2.80***	0.51	2.63***	0.49	2.49***	0.47	2.48***	0.46	2.47***	0.46
Ethnic group (reference: white)														
Not white	1.11	0.15	1.11	0.15	1.13	0.15	1.11	0.15	1.15	0.16	1.16	0.16	1.13	0.15
Country (reference: England and Wales)														
Scotland	0.65***	0.07	0.66***	0.07	0.65***	0.07	0.67***	0.08	0.65***	0.07	0.65***	0.07	0.67***	0.08
Highest level of education (reference: none/other)														
Degree and equivalent	0.38***	0.06	0.39***	0.07	0.39***	0.07	0.37***	0.06	0.40***	0.07	0.40***	0.07	0.42***	0.07
A level and equivalent	0.37***	0.08	0.40***	0.09	0.39***	0.08	0.39***	0.09	0.40***	0.09	0.40***	0.09	0.40***	0.09
O level and CSE	0.97	0.15	0.99	0.15	0.99	0.15	0.99	0.15	0.99	0.15	0.99	0.15	1.00	0.15
Not known	1.77**	0.29	1.84***	0.31	1.74**	0.29	1.83***	0.31	1.74*	0.29	1.73*	0.29	1.63*	0.28
Occupation (NS-Sec) (reference: semi-routine and routine occupations)														
Managerial and professional occupations	0.27***	0.05	0.27***	0.05	0.27***	0.05	0.32***	0.06	0.30***	0.06	0.30***	0.06	0.37***	0.07
Intermediate occupations	0.64*	0.11	0.63*	0.11	0.66*	0.12	0.71†	0.13	0.68*	0.12	0.68*	0.12	0.75	0.14
Small employers and own account workers	0.66	0.17	0.67	0.17	0.68	0.18	0.77	0.20	0.74	0.19	0.74	0.19	0.68	0.18
Lower supervisory and technical occupations	0.50*	0.14	0.51*	0.15	0.51*	0.15	0.58†	0.17	0.54*	0.16	0.54*	0.16	0.61†	0.18
Not classified	1.54**	0.19	1.24	0.16	1.36*	0.17	1.47**	0.19	1.33*	0.17	1.32*	0.17	1.22	0.16
Long-standing illness or disability (reference: no/not known)														
Yes	2.56***	0.27	2.63***	0.28	2.55***	0.27	2.52***	0.27	2.48***	0.27	2.49***	0.27	2.44***	0.27

Table 8 (continued)

Variable	Model1 OR	SE	Model 2a OR	SE	Model 2b OR	SE	Model 2c OR	SE	Model 2d OR	SE	Model 3 OR	SE	Model 4 OR	SE
Unemployed at survey date (survey) (reference: no)														
Yes			3.55***	0.74										
In receipt of JSA and/or IS (administrative data) (reference: no)														
Yes					2.31***	0.40					1.15	0.25		
Months unemployed last 5 years (survey) (reference: 0 months)														
<6 months							1.78**	0.41						
6-11 months							4.02***	1.03						
12+ months							3.48***	0.50						
Months of JSA and/or IS receipt last 5 years (administrative data) (reference: 0 months)														
<6 months									1.81**	0.33	1.77**	0.33	1.91***	0.37
6-11 months									3.01***	0.70	2.91***	0.69	3.06***	0.73
12+ months									2.66***	0.41	2.47***	0.47	2.45***	0.38
Earnings last 5 years (administrative data) (reference: higher tertile)														
0 earnings													2.84***	0.54
Lower tertile													2.70***	0.55
Middle tertile													2.21***	0.44
Intercept	0.15***	0.03	0.15***	0.03	0.15***	0.03	0.12***	0.02	0.14***	0.03	0.1***	0.03	0.06***	0.02
Observations[a]	2,974		2,974		2,974		2,974		2,974		2,974		2,974	
Pseudo R²	0.250		0.261		0.256		0.277		0.267		0.267		0.276	
Log-likelihood	-1,371.79		-1,351.86		-1,359.59		-1,321.14		-1,340.66		-1,340.45		-1,323.48	
Likelihood ratio χ²[b]	912.93***		39.84***		24.39***		101.29***		62.26***		62.67***		96.61***	
Degrees of freedom[b]	23		1		1		3		3		4		6	
AIC[b]	2,791.57		2,753.73		2,769.18		2,696.28		2,735.31		2,736.90		2,706.96	
BIC[b]	2,935.51		2,903.67		2,919.12		2,858.22		2,897.25		2,904.83		2,886.89	

*** $p<0.001$; ** $p<0.01$; * $p<0.05$; † $p<0.10$

[a] The number of observations differs from that presented in Table 6, due to the absence of weighting

[b] Likelihood ratio test for nested models is computed as 2*[(Log-likelihood of unrestricted model (e.g. model 2a)-Log-likelihood of restricted model (e.g. model 1, baseline model))]; The Likelihood ratio χ^2 when comparing Model 3 versus Model 2b is 38.28*** (df=3); Model 3 versus Model 2d is 0.42 (df=1); Model 4 versus Model 2d is 34.36*** (df=3); Akaike Information Criterion (AIC)=−2*ln(likelihood)+2*k, where k=no. of parameters estimated; Bayesian Information Criterion (BIC)=−2*ln(likelihood)+ln(n)*k, where n=no. of observations and k=no. of parameters estimated

References

Administrative Data Taskforce. (2012). *The UK administrative data research network: Improving access for research and policy*. Online document. London: ESRC/MRC/Wellcome Trust. https://esrc.ukri.org/files/research/administrative-data-taskforce-adt/improving-access-for-research-and-policy/. Accessed 12 Dec 2018.

Bailey, N. (2014). *Lower poverty in Scotland: Pinning down the change*. Glasgow: Policy Scotland.

Bailey, N., Barnes, H., Livingston, M., & Mclennan, D. (2013). Understanding neighbourhood population dynamics for neighbourhood effects research: A review of recent evidence and data source developments. In M. van Ham, D. Manley, N. Bailey, L. Simpson, & D. Maclennan (Eds.), *Understanding neighbourhood dynamics: New insights for neighbourhood effects research* (pp. 23–41). Dordrecht: Springer.

Bailey, N., & Bramley, G. (2018). Introduction. In N. Bailey & G. Bramley (Eds.), *Poverty and social exclusion in the UK. The dimensions of disadvantage* (pp. 1–23). Bristol: Policy Press.

Bailey, N., Flint, J., Goodlad, R., Shucksmith, M., Fitzpatrick, S., & Pryce, G. (2003). *Measuring deprivation in Scotland: Developing a long-term strategy*. Edinburgh: Scottish Executive Central Statistics Unit.

Barnes, J., Belsky, J., Broomfield, K. A., & Melhuish, E. (2006). Neighbourhood deprivation, school disorder and academic achievement in primary schools in deprived communities in England. *International Journal of Behavioral Deveevelopment, 30*(2), 127–136.

Barnes, H., Garratt, E., McLennan, D., & Noble, M. (2011). Understanding the worklessness dynamics and characteristics of deprived areas. Department for Work and Pensions Research report no. 779. https://www.gov.uk/government/publications/understanding-the-worklessness-dynamics-and-characteristics-of-deprived-areas-rr779. Accessed 23 Nov 2018.

Bohensky, M. (2016). Bias in data linkage studies. In K. Harron, H. Goldstein, & C. Dibben (Eds.), *Methodological developments in data linkage* (pp. 63–124). Chichester: Wiley.

Boyle, P., Norman, P., & Rees, P. (2004). Changing places: Do changes in the relative deprivation of areas influence limiting long-term illness and mortality among non-migrant people living in non-deprived households? *Social Science and Medicine, 58*(12), 2459–2471.

Bruckmeier, K., Hohmeyer, K., & Scwarz, S. (2018). Welfare receipt misreporting in survey data and its consequences for state dependence estimates: New insights from linked administrative and survey data. *Journal for Labour Market Research, 52*(16), 1–21. https://doi.org/10.1186/s12651-018-0250-z.

Cabinet Office (2012). *Open data white paper: Unleashing the Potential. Cm 8353.* London: HMSO. https://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf. Accessed 23 Nov 2018.

Department for Education (2017). *Pupil premium: funding and accountability for schools*. London: The Stationary Office. https://www.gov.uk/guidance/pupil-premium-information-for-schools-and-alternative-provision-settings#contents. Accessed 23 Nov 2018.

Department for Education (2018). Eligibility for free school meals, the early years pupil premium and the free early education entitlement for two-year-olds under Universal Credit. Government consultation response. https://assets.publishing.service.gov.uk/government/ uploads/system/uploads/attachment_data/file/692644/Government_response_FSM_and_EYentitlements_under_Universal_Credit.pdf. Accessed 23 Nov 2018.

Department for Work and Pensions (2018). *Households below average income: 1994/95-2016/17. London:* Department for Work and Pensions. https://www.gov.uk/government/statistics/households-below-average-income-199495-to-201617. Accessed 23 Nov 2018.

Department of the Environment. (1994). *Index of local conditions: An analysis based on 1991 census data*. London: Department of the Environment.

Department of the Environment, Transport and the Regions (2000). *Indices of deprivation 2000. Regeneration research summary 31, 2000.* London: Department of the Environment, Transport and the Regions.

Dex, S., & McCulloch, A. (1998). The reliability of retrospective unemployment history data. *Work, Employment & Society, 12*(3), 497–509.

Evans, C. E. L., & Harper, C. E. (2009). A history and review of school meal standards in the UK. *Journal of Human Nutrition & Dietetics, 22*(2), 89–99.

Francesconi, M., Sutherland, H., & Zantomio, F. (2011). A comparison of earnings measures from longitudinal and cross-sectional surveys: Evidence from the UK. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 174*(2), 297–326.

Fransham, M. (2017). Income and population dynamics in deprived neighbourhoods: Measuring the poverty turnover rate using administrative data. *Applied Spatial Analysis.* https://doi.org/10.1007/s12061-017-9242-6.

Gambaro, L., Joshi, H., Lupton, R., Fenton, A., & Lennon, M. C. (2016). Developing better measures of neighbourhood characteristics and change for use in studies of residential mobility: A case study of Britain in the early 2000s. *Applied Spatial Analysis*, *9*(4), 569–590.

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L.-C., Smith, P., et al. (2017). GUILD: Guidance for Information about linking data sets. *Journal of Public Health, 40*(1), 191–198.

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

Gordon, D. (2011). *Main PSE UK survey sampling frame. PSE-UK Project working paper methods series No. 21*. Bristol: PSE-UK Project. http://www.poverty.ac.uk/sites/default/files/attachments/WP%20Methods%20No.21%20-%20PSE%20Main%20Survey%20Sampling%20Frame%20%28Gordon%2C%20Oct%202011%29_1.pdf. Accessed 16 Jul 2019.

Gordon, D. (2017). *Producing an 'objective' poverty line in eight easy steps.* Bristol: PSE-UK Project. http://www.poverty.ac.uk/sites/default/files/attachments/Steps-to-producing-the-PSEpoverty-line_Gordon.pdf. Accessed 23 Nov 2018.

Gustafsson, U. (2002). School meals policy: The problem with governing children. *Social Policy Administration, 36*(6), 685–697.

Hand, D. J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Societ.y Series A (Statistics in Society), 181*(3), 555–605.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., et al. (2017a). Challenges in administrative data linkage for research. *Big Data & Society*. https://doi.org/10.1177/2053951717745678.

Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., et al. (2017b). A guide to evaluating linkage quality for the analysis of linked data. *International Journal of Epidemiology, 46*(5), 1699–1710.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis* (2nd ed.). Boca Raton: CRC Press/Taylor and Francis Group.

Hobbs, G., & Vignoles, A. (2010). Is children's free school meal 'eligibility' a good proxy for family income? *British Educational Research Journal, 36*(4), 673–690.

Holterman, S. (1975). Areas of urban deprivation in Great Britain: An analysis of the 1971 census data. *Social Trends, 6,* 33–45.

House of Commons Education Committee (2014). *Underachievement in education by white working class children. First report of session 2014–15.* London: The Stationary Office. https://publications.parliament.uk/pa/cm201415/cmselect/cmeduc/142/142.pdf. Access. 2 Nov 2018.

Ilie, S., Sutherland, A., & Vignoles, A. (2017). Revisiting free school meal eligibility as a proxy for pupil socio-economic deprivation. *British Educational Research Journal, 43*(2), 253–274.

Jarman, B. (1983). Identification of underprivileged areas. *British Medical Journal, 287,* 130–131.

Kho, M. E., Duffett, M., Willison, D. J., Cook, D. J., & Brouwers, M. C. (2009). Written informed consent and selection bias in observational studies using medical records: Systematic review. *British Medical Journal, 338*(b866), 866–873.

Kounali, D., Robinson, T, Goldstein, H., & Lauder, H. (2008). *The probity of free school meals as a proxy measure for disadvantage.* Bristol: Centre for Multilevel Modelling, University of Bristol. http://www.bath.ac.uk/research/harps/Resources/The%20probity%20of%20FSM%20revised%202014.7.08.pdf. Accessed 21 Nov 2018.

Lord, A., Easby, J., & Evans, H. (2013). *Pupils not claiming free school meals—2013*. London: Department for Education. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/266339/DFE-RR319.pdf. Accessed 21 Nov 2018.

Lynn, P., Jackle, A., Jenkins, S. P., & Sala, E. (2012). The impact of questioning method on measurement error in panel survey measures of benefit receipt: Evidence from a validation study. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 175*(1), 289–308.

Maître, C., Whelan, R., & Layte, B. (2003). *Female partner's income contribution to the household income in the European Union.* European Panel Analysis Group Working Paper no. 43. Colchester: Institute for Social and Economic Research, University of Essex. https://www.iser.essex.ac.uk/research/publications/506045. Accessed 23 Nov 2018.

Manley, D. J., van Ham, M., Bailey, N., Simpson, L., & Maclennan, D. (2013). Neighbourhood effects or neighbourhood based problems? A policy context. In D. J. Manley, M. van Ham, N. Bailey, L. Simpson, & D. Maclennan (Eds.), *Neighbourhood effects or neighbourhood based problems? A policy context* (pp. 1–23). Dordrecht: Springer.

McKay, S (2012). Evaluating approaches to Family Resources Survey data linking. Department for Work and Pensions Working Paper 110. https://www.gov.uk/government/ uploads/system/uploads/attachment_data/file/222871/WP110.pdf. Accessed 23 Nov 2018.

Moore, J. C., Smith, P. S. W., & Durrant, G. B. (2018). Correlates of record linkage and estimating risks of non-linkage biases in business data sets. *Journal of the Royal Statistical Society: Series A, 181*(4), 1211–1230.

Noble, M., Wright, G., Smith, G., & Dibben, C. (2006). Measuring Multiple Deprivation at the small-area level. *Environment and Planning A: Economy and Space, 38*(1), 169–185.

Norman, P., Boyle, P., Exeter, D., Feng, Z., & Popham, F. (2011). Rising premature mortality in the UK's persistently deprived areas: Only a Scottish phenomenon? *Social Science and Medicine, 73,* 1575–1584.

Paugam, S., & Russel, H. (2000). The effects of employment precarity and unemployment on social isolation. In D. Gallie & S. Paugam (Eds.), *Welfare regimes and the experience of unemployment in Europe* (pp. 243–264). New York: Oxford University Press.

Rao, J. N. K., & Scott, A. J. (1984). On Chi squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics, 12*(1), 46–60.

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B., & Weir, D. R. (2012). Linking survey and administrative records: Mechanisms of consent. *Sociological Methods & Research, 41*(4), 535–569.

Sala, E., Knies, G., & Burton, J. (2014). Propensity to consent to data linkage: Experimental evidence on the role of three survey design features in a UK longitudinal panel. *International Journal of Social Research Methodology, 17*(5), 455–473.

Smith, G, Noble, M, Anttila, C., Gill, L., Zaidi, A., Wright, G., Dibben, C., & Barnes, H. (2004). The value of linked administrative records for longitudinal analysis. Report to the ESRC National Longitudinal Strategy Committee. Oxford: Department of Social Policy and Social Work, University of Oxford. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.483.6298&rep=rep1&type=pdf. Accessed 23 Nov 2018.

Social Exclusion Unit. (2001). *A new commitment to neighbourhood renewal: National strategy action plan*. London: Social Exclusion Unit, Cabinet Office.

Social Exclusion Unit. (2004). *Jobs and enterprise in deprived areas*. London: Social Exclusion Unit, Cabinet Office.

Strand, S. (2011). The limits of social class in explaining ethnic gaps in educational attainment. *British Educational Research Journal, 37*(2), 197–229.

Townsend, P. (1979). *Poverty in the United Kingdom: A survey of household resources and standards of living*. Berkeley: University of California Press.

Townsend, P. (1987). Deprivation. *Journal of Social Policy, 16*(2), 125–146.

UK Statistics Authority (2017). *Thirteenth meeting of the Administrative Data Research Network Board*. London: UK Statistics Authority. https://www.statisticsauthority.gov.uk/wp-content/uploads/2017/06/ADRN-05-06-2017.pdf. Accessed 23 Nov 2018.

Wichert, L., & Wilke, R. A. (2012). Which factors safeguard employment? An analysis with misclassified German register data. *Journal of the Royal Statistical Society: Series A, 175*(1), 135–151.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistical Neerlandica, 66*(1), 41–63.