

3. SAMPLES AND THE DATA

3.1 Samples

Section 2.4.2 gave a brief outline of the sampling frame. This section describes in more detail how each of the samples was chosen. Note that the word 'sample' is not used in the traditional 'subset of the population' sense (because information about all individuals in the relevant population is available). Rather, it denotes all individuals in the population who meet the criteria set out below.

Analysis samples were defined as follows:

JSA

All individuals:

- (i) who started a JSA spell during 2003
- (ii) aged 16-57 on the start date
- (iii) whose JSA spell did not start during a NDYP, ND25plus, NDfM or Ezones spell
- (iv) who did not have a basic skills language need

IS and IB

All individuals:

- (i) who attended a work-focused interview (WFI) for the relevant benefit during 2003, and this WFI took place no more than six months after the benefit start date
- (ii) aged 16-57 on the WFI date
- (iii) who did not have a basic skills language need

New Deal programmes

All individuals:

- (i) who started a spell of the relevant programme during 2003
- (ii) aged 18-24 (NDYP), 25-57 (ND25plus) or 16-57 (NDLP) on the programme start date
- (iii) who did not have a basic skills language need

Jobcentre Plus

All individuals who belong to one (or more) of the JSA, IS, IB and New Deal programme samples, or:

- (i) who started a spell of NDfM, NDDP, Basic Skills, WBLA, Ezones or Outreach during 2003
- (ii) aged 16-57 on the start date
- (iii) (if the spell was Basic Skills or WBLA) whose Basic Skills or WBLA spell did not start during a NDYP, ND25plus, NDLP, NDfM, NDDP or Ezones spell
- (iv) who did not have a basic skills language need

A number of things should be noted about these definitions:

- The samples are all individual-based, not spell-based. In cases where an individual had more than one spell that qualified them for inclusion in a given sample, only the first such spell (the one starting earliest) is counted, so individuals do not appear in the sample multiple times. This spell is referred to as the 'qualifying spell'.
- An upper age limit was imposed for all samples to avoid including individuals about to retire. The number of individuals affected by this was relatively small (for example, around 3 per cent of individuals starting JSA in 2003 and 4.5 per cent of individuals starting ND25plus in 2003 were excluded by this rule).

- Individuals identified as having a basic skills language need were excluded. This is because those who don't speak English fluently are likely to have significantly different labour market prospects to those who do. Since one might expect this to be more of a problem for ethnic minorities than whites, it would have been near impossible to find comparable white individuals for this group of ethnic minorities.¹
- JSA spells were not permitted to start during a NDYP, ND25plus, NDfM or Ezones spell because it was judged that such spells were more likely to be continuations of previous JSA spells rather than new claims. (Individuals are recorded as having left JSA when they start a New Deal option even though they are still paid an amount equivalent to the value of JSA. If they have not found a job by the time the option has finished, they return to JSA – recorded as a new spell).
- For the Jobcentre Plus sample, Basic Skills and WBLA spells were not permitted to start during a NDYP, ND25plus, NDLP, NDfM, NDDP or Ezones spell because individuals on any of these programmes could be sent on Basic Skills or WBLA. In such cases it was felt that Basic Skills and WBLA spells should be treated as part of NDYP, ND25plus, etc, rather than spells in their own right.
- For IS and IB, analysis focused on individuals with WFIs because (i) these individuals have ethnicity recorded, and (ii) the WFI is a 'treatment' intervention whose effect can be analysed. For new claimants, WFIs usually take place shortly after the start of the IS or IB spell. But since WFIs were introduced across the country starting in 2003 as part of the rollout of Jobcentre Plus, some individuals with a WFI in 2003 had already been receiving IS or IB for a long period of time. To avoid confusing the stock and flow of claimants (see Sections 2.3.2 and 2.4.2), it would have been desirable to require that new spells in 2003 were registered at a rolled-out Jobcentre Plus office (so claimants would have received a WFI within a few weeks). This, however, turned out to be too difficult to implement. Instead, the WFI date was restricted to be during 2003, and no more than six months after the benefit spell start date.²
- The definition of Jobcentre Plus includes all benefits and programmes with an identifiable treatment. This explains why benefits such as Bereavement Benefit, Incapacity Carers Allowance, etc, were excluded. That said, it would have been impossible to include most, if not all, of these benefits because ethnicity was recorded so poorly. It was not possible to replicate the current DWP definition of Jobcentre Plus (which considers all customers with *any* LMS activity) because employment and benefit outcomes were unavailable for customers who had not claimed a benefit or started a programme.
- Given that individuals may start more than one programme or benefit in the space of a year, a decision was made to restrict the Jobcentre Plus sample to include only the first qualifying spell for each individual. For example, if an individual (aged 20) started claiming JSA on 1st February 2003 and then joined NDYP as soon as they were required to (i.e. on 1st August 2003), then they would only be included in the Jobcentre Plus overall sample as a JSA claimant (and not as an NDYP participant). Given that this process was applied equally to both Ethnic Minority and White Jobcentre Plus customers, this should not introduce any bias into the results, and makes the estimation process simpler.

¹ The only exceptions may have been non-British white individuals, for example, those of Eastern European origin, who make up a very small percentage of the white population of interest.

² It should be noted that individuals who claim IS, but who work less than 16 hours per week, are not required to participate in a Work Focused Interview, hence these individuals will be automatically excluded from the analysis.

3.2 The datasets

The samples described above were all selected using the Work and Pensions Longitudinal Study (WPLS) dataset.³ This is a spell-level dataset that contains information about time on benefits and programmes (from DWP records, sometimes called the 'master index') and time in employment (from HMRC records).

The definition of ethnic parity given in Chapter 2 requires that ethnic minority individuals be compared with 'otherwise identical' white individuals. To identify which individuals were otherwise identical, a wide range of background characteristics was required. In fact, the strategy used in this project requires information on all pre-inflow characteristics likely to affect employment and benefit outcomes (see Section 2.5). Clearly one important determinant of employment and benefit outcomes is past labour market involvement, and this can be derived directly from the WPLS (see Section 3.3.2 for more details). Aside from this, the WPLS also includes a small number of useful characteristics, such as ethnicity, date of birth and sex, a partner flag for JSA claimants, and postcode (enabling local area data to be merged in).

Other administrative data sources were used to provide additional background characteristics:

- 100 per cent National Benefits Database (NBD) provided some information for IS and IB claimants about children, partner, disability and illness
- New Deal datasets provided information for NDYP, ND25plus and NDLP participants about disability and marital status

None of the administrative data sources contain reliable information on education or wealth. As these are likely to be important determinants of labour market outcomes, some alternative source was required. In this case, aggregate data available for small geographic areas may be used as a suitable proxy. Information about local labour market conditions was also needed. Both were obtained from the 2001 Census (see Section 3.3.3 for more details).

3.3 Explanatory variables

This section describes the variables that were created using the above data. It was agreed that, as well as an overall ethnic parity estimate for each benefit and programme, analysis should be conducted by ethnicity, sex and region. Subsection 3.3.1 describes these three variables. Subsections 3.3.2-3.3.4 describe variables included in the analysis that are common to all benefits and programmes, while Subsection 3.3.5 discusses those unique to particular groups.

3.3.1 The 'by' variables: ethnicity, sex and region

Estimates of the degree of ethnic parity were split by ethnicity, sex and region. These three variables are described here.

- **Ethnicity**
Ethnicity information in the WPLS allowed 11 separate ethnic groups to be identified:

Black Caribbean

³ The only exception is the exclusion of individuals with a basic skills language need. This was achieved using a dataset provided by DWP derived from Basic Skills administrative data. Individuals were excluded if (variable names are from Basic Skills datasets): SCRRES=5, or if (SCRRES=1,4 and missing ASSDT1 and non-missing REFDATE1). This condition was based on earlier work done by PSI.

Black African
Other Black
Indian
Pakistani
Bangladeshi
Other Asian
Mixed
Chinese
Other Ethnic Group
Unknown

These were aggregated into five higher-level categories:

Black (comprising Black Caribbean, Black African and Other Black)
Asian (comprising Indian, Pakistani, Bangladeshi and Other Asian)
Pakistani and Bangladeshi (note this is a subset of Asian)
Mixed, Chinese and Other
Ethnic Minorities overall (excluding Unknown)

giving a total of 16 different ethnic minority groups.

- **Sex**
Estimates were produced separately for males and females, and then for males and females pooled. In the latter case, a female indicator was included in the models.
- **Region**
The following 78 regions were identified:

Great Britain as a whole
Any of the 272 disadvantaged group wards (this is a single group)
Six cities (Birmingham, Bradford, Leeds, Leicester, London and Manchester), and rest of Great Britain
69 Jobcentre Plus districts

3.3.2 Past labour market history variables

To derive reliable estimates of the ethnic penalty or premium in labour market outcomes, ethnic minority individuals must be compared with *otherwise identical* white individuals. An important dimension in this is past labour market history, particularly in terms of:

- employment history
- benefit history
- past programme participation

Before summary variables could be constructed, however, WPLS employment and benefit data had to be cleaned up considerably. What follows is a brief outline of the steps taken; more details available on request.

For employment spells, the following steps were taken:

- Following DWP advice and internal DWP work, a number of spells were dropped. This includes spells where DWP are unsure whether they have been matched to the right individual, old spells, spells that finish on or before the day they start, spells where the individual was aged less than 10 on the start date, etc.

- A substantial proportion of individuals had multiple spells starting on the same day (often more than two), suggesting that not all ‘old’ spells had been successfully removed as the data were updated. This was certainly the conclusion reached by internal DWP work⁴, which kept only one job start for each individual on any one day. The rules used here differ slightly from this, but the basic principles were similar. Roughly in order, the rules used were:
 - (i) favour spells with certain dates over those with uncertain dates (uncertain means either that only the tax year in which the spell started and/or ended is known, or that the spell is ongoing)
 - (ii) favour spells from more recent extracts of the data
 - (iii) favour longer spells
- Some spells in the employment data actually were spells on taxable benefits. Many of these were flagged (and therefore easily removed), but this wasn’t possible for all such spells. A set of rules, based on matching benefit and employment dates, was devised to try to eliminate more of these non-employment spells.
- Having carried out all this cleaning, a large number of employment spells still had uncertain start or end dates. For these spells, start dates were set as early as possible and end dates as late as possible, and then an attempt made to reduce the uncertainty using the fact that some benefits are very unlikely to overlap with employment

A number of decisions, in particular those to favour longer spells when deleting duplicated spells and to set uncertain dates to maximise the length of the spell, are likely to have led to the length of employment spells being overstated. But this only matters to the extent that the spells of ethnic minorities and whites are *differentially* affected.

Benefit spells, in contrast, were relatively straightforward to clean up. In short:

- Following DWP advice, a number of spells were dropped (spells with negative duration, spells where the individual was aged less than 10 on the start date).
- The end date of most benefit spells (not JSA) is only known within a two- or six-week window. For consistency with employment spells, the end date for these benefit spells was set to the last possible date the spell could have finished (i.e. the end of the window of uncertainty).

Again, choosing the latest possible end dates is likely to overstate the length of benefit spells, but this only matters only if ethnic minorities and whites are differentially affected.

In the raw data, all employment, benefit and programme information appears in the form of spells (start and end dates – see above). Because of the complexity of spell patterns, this had to be summarised in some way to enable straightforward comparison across individuals. The decisions made about how to do this are now documented.

- Three years’ worth of labour market history was used. Given the 2003 inflow window, the earliest spell information used relates to January 2000, consistent with DWP’s concern that some spells prior to August 1999 may be unreliable.
- For benefit and employment histories, it made logical and practical sense to use the same variable definitions as used for benefit and employment outcomes. These were discussed in Section 2.4.4, but the key features were:

⁴ This work was part of the DWP’s evaluation of the Job Outcome Target pilots. See <http://www.dwp.gov.uk/asd/asd5/rports2005-2006/rrep316.asp>.

- (i) they were defined in *relative* terms, counting from the day the individual started the programme that qualified them to be included in the analysis group (called the time of inflow, $t=0$)
 - (ii) the measures were monthly (30-day intervals)
 - (iii) an individual was defined as being employed (on benefit) in a given month if he or she was employed (on benefit) for 15 or more days in that month; an individual was defined as being sustainably employed if he or she had been continuously employed for at least 3 months
 - (iv) the definition of 'on benefit' included: JSA, NDYP, ND25plus, Ezones, Basic Skills, WBLA, IS and IB⁵
- Monthly employment, sustained employment and benefit variables were created along these lines for each of the six months prior to inflow. This was because previous research has shown that recent labour market history is more important than earlier labour market history in determining future labour market outcomes.
 - One real difficulty with using WPLS employment and benefit information to compare ethnic minorities and whites is that recent immigrants will appear as having no employment or benefit history. If ethnic minorities are more likely to be recent immigrants than whites, there is a real danger that ethnic minorities with a full (but unknown) work history will be compared with whites who have never worked. To try to counter this problem:
 - (i) dummy variables were defined for each of the six months prior to inflow, indicating whether the individual had yet appeared in the WPLS
 - (ii) all 'proportion of time' variables (see below) were calculated relative to time since first appearing in the WPLS
 This issue is discussed in more detail in Section 4.3.4.
 - Labour market history in the three years prior to inflow was also summarised more broadly using two variables: proportion of time employed in these months (or since first appearing in the WPLS, whichever is more recent), and proportion of time on benefit in these months (or since first appearing in the WPLS, whichever is more recent).⁶
 - NDYP and ND25plus are mandatory after 6 months or 18 months of JSA receipt, respectively. This raises the question of whether, for these groups, earlier labour market history might be more informative than more recent history because many individuals have been unemployed and on JSA in the run-up to inflow. It was decided, however, not to use substantially different history variables. This was both because a substantial proportion of individuals enter these programmes early (so won't necessarily have been unemployed and on JSA for all of the recent past), and because summarising earlier history in a more detailed way seemed to make little difference to the results. The only concession was the inclusion of an indicator of early entrant status.
 - It was thought that individuals with a history of inactive benefit receipt were likely to have considerably different employment and benefit prospects to those who had claimed only active benefits. Therefore, two additional dummy variables were included: one flagging individuals who had spent some, but not more than 50 per cent, of their time on inactive

⁵ "This definition is similar to that used in "The Longer Term Impact of the New Deal for Young People" by Ian Beale, Claire Bloss and Andrew Thomas (April, 2005)

⁶ It had been thought that two sets of variables might be necessary: one for large subgroups that contained dummy variables for months 7-36 (like the variables for months 1-6), and one for small subgroups summarising the information more parsimoniously (as described above). It turned out, however, that the choice made little difference to the results, so only the parsimonious summary was used.

benefits since first appearing in the WPLS, and one flagging individuals who had spent more than 50 per cent of their time on inactive benefits.

- In the WPLS there are a substantial number of employment spells lasting just one day. HMRC advice suggests these are often cases where the end date is known for certain, but the start date is unknown. To avoid discarding this information completely, a dummy variable was created for individuals with at least one such spell in the three years prior to inflow, indicating that they had been employed for at least part of this period.
- Past participation in various programmes is likely to reveal information about individuals. In particular, participation in voluntary programmes may suggest a willingness to be proactive in improving employment prospects, whilst participation in Basic Skills may indicate individuals who are less likely to be successful in the labour market. In line with this, indicators of voluntary programme participation, Basic Skills need and Basic Skills attendance during the three-year history period were created.

3.3.3 Individual characteristics

Aside from labour market history, there are a number of other individual characteristics it is important to control for. These include:

- gender (only for estimates with males and females pooled)
- age
- month of inflow
- educational qualifications
- wealth

Of these, the last three merit brief discussion.

There are pronounced seasonal patterns in employment and benefit receipt so it is important to control for the month of inflow onto the benefit or programme of interest.

Education is an important determinant of labour market outcomes. Unfortunately, however, none of the administrative datasets used for this project contained reliable education information.⁷ To address this problem, local area census data was used as a proxy. The lowest level of aggregation at which ONS were willing to provide education for unemployed or inactive people (the relevant population of interest for benefit claimants and New Deal participants)⁸ split by white/non-white status⁹ was Super Output Area (SOA, around 750 households). This was used to calculate the proportion of unemployed or inactive individuals in each combination of SOA and white/non-white status with:

- no qualifications (this is the omitted category)
- level 1 qualifications
- level 2 qualifications
- level 3 qualifications
- level 4 or 5 qualifications
- unknown qualifications

⁷ The New Deal datasets included variables intended to record highest qualification, but the information was missing for the vast majority of the sample, so could not be used.

⁸ For JSA, ND25plus, NDYP and NDLP, unemployed individuals were thought to be the most appropriate group. For IS and IB, a combination of unemployed and inactive individuals was used.

⁹ ONS were unwilling to provide information for more disaggregated ethnic groups.

For each white/non-white individual in a given SOA, the relevant information was mapped in using home postcode. In this context, the information can be interpreted as a probability that the individual has each level of qualification.

There are two potential problems with this approach. First, there is no guarantee that the correlation between actual qualifications and the SOA-level proxy is strong, and without data on actual qualifications, it is not something that can be tested. But given the absence of an alternative, there was little choice but to use this measure. The second problem relates to the fact that the Census was collected during 2001, whereas inflow took place during 2003. This means that, not only might the information not be the most up-to-date available (individuals may have gained higher qualifications after the census, but prior to inflow)¹⁰, but it is also possible that individuals in one of the analysis samples lived in an area in 2003 where no ethnic minorities lived when the census was collected (in 2001), meaning that there is no SOA level education information available for them. In fact, this latter issue turned out not to be a problem.

The administrative data also contained little indication of individual wealth, so, as with education, the 2001 census was used to provide a proxy. A number of alternatives were available.¹¹ The one chosen here was the proportion of each ethnic group¹² living in council or other social-rented housing at ward level.¹³

3.3.4 Local labour market characteristics

Employment opportunities and local conditions can vary widely across neighbourhoods. It is therefore important to control for these differences. The variable used to control for these differences was unemployment by travel-to-work area from the 2001 Census. The same concerns about timing outlined above also apply here.

3.3.5 Other variables not common across groups

Some information was only available (and perhaps even only relevant) for particular analysis groups. This Subsection lists the remaining variables that were included in the model, and the programmes and benefits for which they existed.

- Partner (available for: IS, JSA, NDYP, ND25plus; not relevant for: NDLP)
Partner information relates in general to claims for a partner.
- Number of children (available for: IS, NDLP, IB; proxied: JSA, JCP, ND25plus, NDYP)
Information about children relates in general to claims for children. Proxies were constructed using 2001 census data in a way identical to that described above for wealth: the variable measures the proportion of the relevant ethnic group aged under 16 in the relevant ward.

¹⁰ Although the most important thing is that the information was collected prior to inflow.

¹¹ Results did not seem to vary with the measure chosen.

¹² For England and Wales, data was available for: White British, Irish, Other White (combined to provide information for White individuals), White and Black Caribbean, White and Black African, White and Asian, Other Mixed (combined to provide information for individuals of Mixed ethnic origin), Indian, Pakistani, Bangladeshi, Other Asian, Black Caribbean, Black African, Other Black, Chinese and Other Ethnic Groups (each of whom appeared separately in the analysis). For Scotland, data was available for: White, Indian, Pakistani and Other South Asian (this was mapped to Pakistani, Bangladeshi and Other Asian individuals), Chinese and Other (where Other was mapped to Black Caribbean, Black African, Other Black and Other Ethnic Groups).

¹³ The same variable at OA level (approximately 150 households) was originally mapped in, but was missing for a large proportion of individuals (indicating that people had moved across OAs between the census and the time of inflow). Aggregation to SOA level did not seem to make any difference. It was thought that disaggregation by ethnicity was more important than disaggregation by region, so a ward level measure was adopted.

- Age of youngest child (available for: IS, NDLP)
Information about children for IS customers relates to claims for children.
- Disability (available for: IS, NDYP, ND25plus, NDLP; not relevant for: IB)
IS disability information in the NBD relates to receipt of IS disability premium. The majority of individuals are recorded as receiving the premium only after their IS claim had started. But given the possibility of administrative delays, the distinction between whether or not the premium was received from the start of the claim was ignored. New Deal disability information did not seem to be tied to receipt of a premium.
- On IB at inflow (available for: all benefits/programmes; not relevant for: IB)
Being on IB is an indicator of incapacity for work. Therefore it is sensible to control for it in programmes and benefits other than IB.
- On IS at inflow (available for: all benefits/programmes; not relevant for: IS)
Being on IS is an indicator of low income, which may affect your labour market (participation) decisions. Therefore it is sensible to control for it in programmes and benefits other than IS.
- Amount of benefit at inflow (available for: IS, IB)
The purpose of including this information was twofold: as an indication of the amount that might be lost on entering work, and as a proxy for characteristics not available elsewhere. Because there are only a small number of different IB rates, it was possible to classify the amount received into one of the following five categories: (i) £0, (ii) lower rate, (iii) lower rate plus adult dependent payment¹⁴, (iv) higher rate, and (v) other. A similar classification was not possible for IS: here the distribution of amount received was split three ways: (i) less than £40, (ii) at least £40 and less than £60, and (iii) £60 or more.¹⁵
- IB illness information (available for: IB; not relevant elsewhere)
Individuals were classified into 5 categories according to their IB illness code using the International Classification of Diseases.
- Occupation (available for: JSA)
For JSA, the WPLS contains information both about the usual occupation of the individual and the occupation of the job they are seeking. This was used to create a 10-way classification of usual occupations, and indicators of whether the individual was seeking a job belonging to a higher or lower occupational group.
- Previous New Deal spells (available for: NDYP, ND25plus, NDLP)
The number of previous spells on the relevant New Deal programme was included.
- Early entrant (available for: NDYP, ND25plus; not relevant elsewhere)
This indicates whether the participant entered the relevant programme early (i.e. before the qualifying conditions based on JSA receipt had been met).
- Programme indicators (available for: JCP; not relevant elsewhere)

¹⁴ It was hoped that this might provide some information about partners (which is not available elsewhere for IB claimants).

¹⁵ The amount of IS received may provide information about an individual's household income, although because the personal allowance depends on circumstances, for example, whether an individual has a dependent partner or dependent children, it is not necessarily a monotonic relationship.

Since the JCP analysis group pooled individuals on different programmes and benefits, it was necessary to create indicators of the type of programme or benefit that the qualifying spell related to.¹⁶

3.4 Limitations of the data

Some limitations of the data used have already been touched on in previous sections. Here, the main issues are collected together.

- Employment spells for individuals on low earnings may not appear in the WPLS. Employment information in the WPLS is derived from P45 and P46 forms sent to HMRC by employers. It is only compulsory, however, to submit forms for employees earning enough to be subject to income tax. Although some employers submit forms for all employees regardless of their earnings, this is not always true. Therefore, individuals earning below the income tax threshold may appear as having no employment spells. This causes problems if it differs by ethnicity – which it might do if, say, ethnic minorities more often work for small employers who are less likely to submit forms for employees below the income tax threshold. There is no way of telling how much of a problem this is.
- The start and end dates for many employment spells are not known with certainty. Most commonly this manifests itself as 5th April start dates and 6th April end dates (indicating that the tax year in which the employment spell started or finished is known, but not the precise date). There are also a considerable number of spells lasting just one day (usually indicating that HMRC received notification of an end date for a job they didn't know existed). These unknown dates create wide windows of uncertainty of when individuals were actually employed. Again, they pose a problem for this project if unknown dates affect ethnic minority employment histories and/or outcomes differently to white employment histories and/or outcomes.
- Many individuals have multiple employment spells that seem to relate to the same job (for example, they start on the same day and there are too many for them to all be different jobs). If all these spells had the same start and end dates it would not be a problem for employment histories and outcomes as defined in this project (because no account is taken of the number of jobs held). But this is not the case: often the end dates of spells differ (some are certain end dates, others are uncertain within a tax year, others are open spells, etc).
- The end dates for most benefit spells are only known to within a given window (usually two or six weeks). In the raw data the actual end date is randomly set to be some time during that window.
- There is no record of how long individuals have been living in the UK. In a project comparing ethnic minorities and whites, immigration is likely to be a considerable issue. An individual who has just arrived in the UK and starts claiming benefits or joins a programme will never previously have appeared in the WPLS (i.e. they will be recorded as never having previously been employed or on benefit), whereas they may, in reality, have been employed or on benefit in another country. This means that they will be compared to white individuals who have also only just appeared in the WPLS, but who may have been absent for potentially different reasons, for example, studying or starting a

¹⁶ Note that individuals for whom the qualifying spell was NDfM, NDDP, Basic Skills, WBLA, Ezones or Outreach were grouped into an "Other" programme type.

family. The underlying assumption in this project must be that Ethnic Minorities and Whites are absent from the WPLS for the same reasons.

- The range of background characteristics available in the administrative data is limited. In particular, there is no reliable education information, and (in the version of the data used for this project), no indication of income or wealth. As described above, it is possible to use census data to proxy for this missing information, but this is likely to fall some way short of having the information for the individuals themselves. This is because the proxy is an area-level average by ethnicity. Where many individuals of a given ethnicity live in a particular area, the proxy is completely reliant on individuals in that area having similar qualifications to the individual being proxied.
- Recent migrants to the UK may not speak English very well and are therefore likely to find it much harder to find work than an otherwise equivalent individual. The only source of information about language needs is the ESOL (English for Speakers of Other Languages) dataset. This can help identify individuals who don't speak English, but it isn't clear that all such individuals will be captured (and therefore dropped from the analysis).