Social Mobility Commission

Background to the research informing the employer toolkits

Introduction

In February 2020, the Social Mobility Commission (SMC) published a toolkit aimed at encouraging and enabling employers in all industry sectors to increase levels of socio-economic diversity and inclusion within their organisations. The toolkit emphasises the business case for increasing inclusion and diversity, describes a strategic approach involving six elements, and provides guidance on steps that employers can take to move forwards in each of those elements.

Recognising that the contexts and challenges faced in different industry sectors vary, the SMC is also producing sector-specific toolkits focussing on financial and professional services (FPS), the creative industries, the public sector and retail. The work to inform these toolkits was conducted over the course of several months starting from March 2020. The FPS toolkit has already been published; the others are scheduled for 2021. Secondary quantitative analysis of data from the ONS's Labour Force Survey (LFS) and primary qualitative research were conducted to inform these sector-specific toolkits. This background paper is intended to provide an overview of the data sources used and techniques employed.

The quantitative analysis involved extensive secondary analysis of data from the Labour Force Survey to provide the majority of all descriptive statistics pertaining to the workforce of each sector. In addition, the Inter-Departmental Business Register (IDBR) was consulted to provide all descriptive statistics pertaining to the businesses in each sector. Details of the other data sources referenced in the production of the descriptive statistics can be found in the 'Data Sources' section below. Regression modelling was also employed for each of the main priority sectors to investigate whether socio-economic background (SEB) is associated with pay once possible confounders have been controlled for.

The qualitative research involved conducting short telephone interviews with diversity and inclusion leads and programme managers at businesses and organisations working on social mobility. Where possible, leads were identified from among the top 75 employers in the 2019 Social Mobility Employer Index.[1] Recruitment was further complimented by employer leads previously known to the SMC as well as employer referrals. A grand total of 32 interviews were conducted with employers from across the four toolkit industries. Most interviews were on social mobility programmes and interventions, and two were personal case studies. The intervention interviews served two purposes: to collect case studies on social mobility best practice and to understand the wider business rationale and approach to working on socio-economic diversity. The personal case study interviews focused on the life and career journeys of individuals from disadvantaged backgrounds who have succeeded into leadership roles. These personal case

[1] The Social Mobility Index ranks employers on the steps they are taking to improve representation from all backgrounds. www.socialmobility.org.uk/index

studies seek to acknowledge the specific socio-economic barriers to career entry and progression and recognise the key factors that can help individuals overcome these.

The overall quantiativeresearch findings helped inform the toolkits with sector-specific insights, while the case studies have served to showcase best practice and inspire employers to improve their social mobility agenda. Case studies can be found in the published toolkits as well as on the SMC case study microsite.

Data sources

The table below provides a summary of the data sources used to inform the employer toolkits:

Data source	Summary of suitability	Accessibility	Usage in the research
Annual Survey of Hours and Earnings (ASHE)	Business survey covering 1% of the workforce. Employers provide information on hours and earnings for sampled employees. It is deemed to be more robust than the LFS (as LFS pay data is self-reported).	Lots of published aggregated data by 4-digit SIC code.	All gender-pay gap estimates provided in the desk research were sourced from the 2019 ASHE. Despite being deemed a more reliable source of pay data, the ASHE was not used in the modelling of the "class-pay gap". The ASHE was not suitable for this analysis, as it does not capture information on socio-economic background and many other demographic variables.
Business Population Estimates (BPE)	Statistics on businesses from the Department for Business, Energy and Industrial Strategy (BEIS). Uses the LFS as well as the IDBR to produce estimates that include *unregistered businesses*.	Published data is not particularly granular. Priority sectors would be hard to identify.	Referenced to provide the estimates of the overall workforce size for small and medium-sized businesses, as the LFS only provides information on small and medium sized *workplaces.*[2] However, the LFS was still used to produce all the workforce statistics for SMEs, as the BPE does not provide detailed publicly available demographic information.
Inter-Departmental Business Register (IDBR)	Covers *all* registered businesses and provides information on location, ownership, sector and size.	Publicly available at a granular level through Nomis (a service giving free access to UK labour market	All business level analysis in the desk research was conducted using the data from

[2] It is important to note that all the analysis of workforce in this research is pertaining to the workforce of small and medium workplaces rather than enterprises. This is because analysing workforce requires the use of the Labour Force Survey (LFS), which only asks about a respondent's workplace rather than the enterprise they work for. Therefore, the research only provides an indication of the characteristics of the workforce of small and medium enterprises and should be treated with caution.

		statistics from official sources).	the 2019 IDBR on enterprise groups available on Nomis.[3]
Labour Force Survey (LFS)	Large survey of the UK population - used for official Labour Market estimates. Captures detailed information on demographics, work, pay and qualifications. The large sample size means it is the most appropriate data source for conducting analysis of priority industries.	Micro data can be accessed through UK Data Service. Aggregated data can be accessed through NOMIS	All the workforce statistics (bar the gender-pay gap estimates) in the desk research were generated using data from the LFS. All the data used in the modelling of the "class-pay gap" was also sourced from the LFS, as it provided information on key demographics.

All analysis using LFS data was carried out on those aged 16 and over in England and whose main job is in one of the priority sectors of interest.[4]

For all research questions pertaining to the **socio-economic background** of the workforce, the results have been obtained using data from the July - September quarters of the LFS from the years 2017 to 2019. This quarter was selected as it is the only quarter of the LFS survey that includes questions on socio-economic background. Three quarters were pooled together to create a sufficient sample size for the analysis.

For the research question related to **training offered** to the workforce, the results have been obtained using data from the January - March quarters of the LFS from the years 2017 to 2019. This quarter was selected as it is the only quarter of the LFS survey that includes questions on training offered to workers, in addition to training completed or currently ongoing. Three quarters were pooled together to create a sufficient sample size for the analysis.

For all other research questions pertaining to workforce (apart from the gender-pay gap analysis), the results have been obtained using data from all four quarters of the LFS from 2019. All four quarters were used as this better reflects the profile of workers over an entire year - including any seasonal employment. 2019 data was used as it is the most recent data available at the time of analysis.

[3] Information accompanying the IDBR states: "an enterprise can be thought of as the overall business, made up of all the individual sites or workplaces. It is defined as the smallest combination of legal units (generally based on VAT and/or PAYE records) that has a certain degree of autonomy within an enterprise group". For more information refer to the UK Business: Activity, Size and Location section on the ONS website.

[4] Studies carried out on a different cohort or at an alternative point in time may yield different results.

Sample sizes

Analysis of the workforce was only provided for cohorts where there was a minimum sample size of approximately 100 respondents available. This is the generally accepted threshold for reliable statistical reporting on a population. Moreover, this practice also ensures that the risk of statistical disclosure arising from small cell counts is minimised.

The LFS sample sizes available for the desk research are shown in the table below:

Priority sector	LFS Jul-Sep quarter 2017-2019 Respondents	LFS 2019 all quarters Respondents	LFS Jan-Mar quarter 2017-2019 Respondents
Financial and Professional services	**5,124**	**6,813**	**5,190**
Financial services sub-sector	1,675	2,143	1,706
Accounting sub-sector	989	1,265	981
Management consultancies sub-sector	1,391	1,967	1,401
Legal sub-sector	782	1,039	811
Creative industries	**6,197**	**8,228**	**6,256**
Creative film sub-sector	687	933	732
Creative publishing sub-sector	613	790	598
Public sector	**21,828**	**28,175**	**22,296**
Retail	**9,345**	**11,522**	**9,795**
Small workplaces	**55,828**	**72,000**	**57,216**
Medium workplaces	**20,602**	**26,773**	**21,030**

Sample sizes are further reduced where there is missingness in the data (also see page 5).

Only data from the 2017, 2018 and 2019 July-September quarters of the LFS was used in the modelling, as the socio-economic background questions are only included in this quarter of the LFS each year.

The socio-economic background variable

Socio-economic background (SEB) has been defined, for the purposes of this research, as the occupation of the survey respondent's main wage earner when they were 14.[5]

In addition, the SMC also requested that respondents whose main wage earners were not working when they were 14 were also included in the routine and manual occupations SEB category. This was to capture the effects of long-term scarring caused by stretches of unemployment, as a form of extreme disadvantage. The analysis presented in the desk research was trialled both including

[5] This corresponds to the variable SMSOC103 in the LFS.

and excluding these respondents, as well as those who were not living at home when they were 14; all these variations did very little to affect the results of the research.

Creating the NS-SEC SEB categories

The LFS only provides standard occupational classification (SOC) categories for SEB information, as opposed to NS-SEC occupational categories. The SMC requested that these categories were transformed into NS-SEC categories, to be in line with other published research. This meant that SEB SOC categories needed to be mapped to NS-SEC categories.

The NS-SEC category of the main wage earner when the survey respondent was 14 was generated by mapping the 3-digit SOC categories of the main wage earner in the LFS to NS-SEC categories. This was done using the existing mapping of 4-digit SOC codes to NS-SEC categories by the ONS.[6] Where not all 4-digit SOC codes within a 3-digit SOC category mapped to the same NS-SEC category, the most common NS-SEC category for that 3-digit SOC code was chosen.

Unfortunately, 4-digit SOC codes were unavailable to us in the publicly available LFS data, which meant the mapping had to be done using 3-digit SOC codes. This means there may be small discrepancies with other analyses using 4-digit SOC codes.

Re-weighting the data to account for SEB missingness

There is a relatively high incidence of missingness in the SEB data, originating from respondents from previous quarters of the LFS being included in the July-September quarter datasets. The existing LFS weighting did not account for this, so a separate weight was generated and applied to the data to help remedy any resulting observable bias.[7]

To provide a comparison for some of the key descriptive statistics produced in each of the sector toolkits, we additionally produced the same descriptive statistics for the general population. However, whilst we generated separate weights for the sector analysis to remedy any observable bias, doing the same for the general population figures was outside the scope of this work. As a result, there may be discrepancies with other published analysis on general population figures using the LFS.[8]

[6] https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2020

[7] Variables included in the non-response model were: age, gender, ethnicity, region, number of dependents, full time/part time status, household size, nationality, highest qualification, NS-SEC category and disability status.

[8] For example, these statistics on the general population may differ from those provided as reference in the toolkits.

Modelling the “class-pay gap”: approach & caveats

We conducted regression modelling on each of the main priority sectors to investigate whether socio-economic background (SEB) is associated with pay once possible confounders have been controlled for. The log transformation of self-reported weekly pay was used as the outcome variable for this modelling and SEB was taken to be the NS-SEC category of the main wage earner in the respondent’s household when they were aged 14.[9] This modelling allows us to compare pay across individuals with different socio-economic backgrounds but who are otherwise similar, thus providing an estimate of the “class pay gap”.

The predictor variables used in the modelling are summarised in the table below. The rationale behind this selection of variables was to ensure our modelling work was consistent with that done by Sam Friedman in his book *The Class Ceiling.* Certain variables listed in this work were unavailable in the July-September quarter of the LFS or suffered from a high incidence of missingness, and were therefore not included in the model specification. Nationality was also excluded from the model as it was found to be colinear with ethnicity.

Category of predictors	Variables included
Personal characteristics	Gender, age, ethnicity, disability status, health status, marital status, socio-economic background indicator
Geography	Region of workplace
Household characteristics	Number of dependents in household
Employment	NS-SEC category, length of time in current employment, hours worked per week, working from home status (excluded from the PFS and Retail models due to sample size restrictions), training indicator, size of workforce, sector (SME model only), type of public organisation (Public model only)
Education	Highest qualification obtained
Other controls	Year of LFS data collection.

Sample sizes for the modelling were restricted to respondents with socio-economic background data and pay data.[10] Furthermore, the longitudinal design of the LFS further restricted sample sizes.[11] The final LFS sample sizes available for the modelling can be seen in the table below:

[9] As mentioned previously, this data is likely to be less reliable than other sources of pay data. The ONS recommends that the Annual Survey of Hours and Earnings (ASHE) should be used for robust analysis of earnings, as it is collected directly from employers. However, the ASHE was not suitable for this analysis, as it does not capture information on SEB and many other demographic variables.

[10] The pay questions are only asked to c.40% of the LFS sample each quarter (waves 1 and 5). Furthermore, these questions are only asked to employees (not the self-employed). Finally, some respondents refuse to provide earnings information (c.20%) and will therefore also be missing from the modelling.

[11] The LFS design is longitudinal - with selected addresses asked to participate in five quarterly waves of the survey. This means that in each annual July-September quarter, roughly half of the sample providing pay information will have also participated in the previous year’s July-September survey.

Priority sector	Sample size
Retail	1,652
Professional and Financial Services	837
Creative Industries	895
Public Sector	4,656
Small and Medium-sized workplaces	12,600

There are various forms of bias that may be present in the models, which could impede the identification of the true association of socio-economic background with pay.

Selection bias

Respondents with pay and SEB data may be systematically different to those who refused to answer these questions or were missing this data in a way that is associated with earnings. This would mean that the estimated relationship between socio-economic background and pay that we observe is specific to the type of respondents that chose to answer the questions on earnings and not to the entire population.

Omitted variable bias & other forms of model misspecification

Omitting a variable that is correlated with both SEB and the outcome of interest (pay) could cause the model estimates to be biased. While the LFS includes a wide range of demographic variables that can be included in the model, it does not fully capture many individual specific characteristics such as ability, motivation and experience. Moreover, even though we are controlling for various employment characteristics, these only capture broad differences in job roles and neglect to capture detailed features such as seniority and specific sub-sectors effects. It is also possible that the relationship between pay and the included model predictors may be specified incorrectly. For example, there may be important non-linearities in the relationships between pay and the model predictors. We attempted to mitigate this risk by running diagnostic tests on the models and fitting different specifications to test how sensitive the findings are to these model choices.

Causal inference issues

It is important to note that we are most likely not capturing the entire effect of SEB on pay. The models isolate the direct influence of SEB on pay; but in addition to this, SEB is likely to have affected variables which in turn have their own influence (for example, the respondent's NS-SEC and highest qualification obtained). This 'indirect' influence of SEB is not reported. As a result, it is not fair to say that the models estimate the *total impact* of SEB on pay. Instead, we can use this analysis to estimate the *average difference* in pay for people who have different levels of SEB but are similar in other respects.

Variable definitions

Most of the variables included in the modelling and desk research are self-explanatory, but we have provided a more detailed explanation of some of the more obscurely named variables in the table below.

Variable name	Where found	Details	Codes in LFS
Training status	Modelling	Indicator of whether a respondent has received job related training or education in the last 3 months.	ED13WK
Health indicator	Modelling	Indicator of whether a respondent has reported having health conditions/illnesses lasting 12 months or more.	LNGLST
Number of dependents	Modelling	The number of dependent children in the household aged under 19.	HDPCH19
Workforce size	Modelling	Estimated number of employees at the respondent's workplace. Small: 0 to 49 employees Medium: 50 – 249 employees Large: 250+ employees	MPNR02 & SOLOR
Socio-economic background	Desk research	Socio-economic background is defined as previously outlined. However, in some instances in the desk research, a socio-economic background category named “Working class occupations” is referenced. This category covers both the “Intermediate occupations” and “Routine and manual occupations” SEB categories. This category was created to enable analysis where small sample sizes would have otherwise meant that analysis by SEB was not possible.	SMSOC103