Department for Business Innovation & Skills

RESEARCH

BIS RESEARCH PAPER NUMBER 81C

2011 Skills for Life Survey: Small Area Estimation Technical Report

DECEMBER 2012

The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Business, Innovation and Skills.

Department for Business, Innovation and Skills

1 Victoria Street

London SW1H 0ET

www.bis.gov.uk

Research paper number 81C

November 2012

Authors and Acknowledgements

Authors

Alex Gibson is Principal Researcher and Director of RAE Consulting, and Innovation and Research Fellow in the School of Health, Education and Society at the University of Plymouth.

Paul Hewson is Associate Professor in Statistics in the School of Computing and Mathematics at the University of Plymouth.

Acknowledgements

We are grateful for guidance and feedback from Carrie Harding and Joel Williams of TNS/BMRB and Kristopher Chapman and Ashley Buckner of the Department for Business, Innovation and Skills. We also acknowledge access to the following sources:

2001 Census Data: Obtained using Casweb (http://casweb.mimas.ac.uk/), a census interface provided by the Census Dissemination Unit, Mimas (University of Manchester). [Accessed 12/1/2012]. Source: 2001 Census: Standard Area Statistics (England and Wales). Census output is Crown copyright and is reproduced with the permission of the Controller of HMSO.

Benefit Data: LSOA-level Department of Work and Pensions 4th Quarter data on the number of people claiming Incapacity Benefit and Severe Disablement Allowance has been extracted from the Work and Pensions Longitudinal Study (WPLS) dataset available via NOMIS (http://www.nomisweb.co.uk/). [Accessed 12/1/2012.] ONS Crown Copyright Reserved [from Nomis on 15 July 2011].

Postcode Directory: *Office for National Statistics Postcode Directory (ONSPD)* Open February 2011 edition, 2011. Available at http://www.sharegeo.ac.uk/handle/10672/203. [Accessed 12/1/2012.] Contains National Statistics data ©Crown Copyright and database right 2011.

Boundary data: MSOA and Local Authority Boundary data has been obtained via UKBorders (http://edina.ac.uk/ukborders/). [Accessed 12/1/2012.] This data is provided with the support of the ESRC and JISC and uses boundary material which is copyright of the Crown, the Post Office and the ED-LINE consortium. Contains: ONS, Super Output Area Boundaries. Crown copyright 2004; and Ordnance Survey data © Crown Copyright and database right 2010.

Contents

Research Objectives

1. The project's principal objectives were to apply Small Area Estimation methods to the *2011 Skills for Life (SfL) Survey* in order to generate local area estimates of the number and proportion of adults living in households with defined literacy, numeracy and ICT skills, as well as the number and proportion of adults in households who do not speak English as a first language. The resulting database of local area estimates is accompanied by a *Technical Report* (this document) and a shorter *User Guide*.

2. Small Area Estimation (SAE) has been implemented for ***Middle Layer Super Output Areas (MSOAs)*** and attributed to other geographies on the basis of addresses listed in the February 2011 *Open National Statistics Postcode Directory* (*ONSPD*)[1]. Thus, in addition to the core set of modelled MSOA-level estimates (n=6,781), local area estimates have been proportionally attributed, as requested by the project brief, to 7,932 ***Standard Table (ST) wards*** (which are precisely equivalent to the 2003 Statistical Wards used to report small area estimates derived from the *2003 Skills for Life Survey*[2]), 7,972 ***2005 Statistical wards*** and 7,618 ***2011 Council wards***, as well as to ***2011 Parliamentary Constituencies*** (n=533), ***Local Authorities*** (n=326) and ***Local Enterprise Partnership areas*** (n=37).[3] Details on each of these geographies is provided in the Glossary (Section 0).

3. The MSOA-level estimates have been generated using fully Bayesian hierarchical Generalised Linear Mixed Models (GLMM) which link outcome data from the *2011 SfL Survey* with local area covariate data drawn from the *2001 Census* and the Department of Work and Pensions' (DWP) *Work and Pensions Longitudinal Study* dataset[4]. The approach

---

1 ONS, *Office for National Statistics Postcode Directory (ONSPD)* Open February 2011 edition, 2011. Available at http://www.sharegeo.ac.uk/handle/10672/203. [Accessed 12/1/2012.]

2 Gibson, A., Bailey, T, and Fraser, D. (2004) *Demographic mapping of the 2003 Skills for Life Survey to local areas*. Technical Report for the Department for Education and Skills, December 2004. We advise great caution comparing the two sets of estimates. Not only are the population denominators different, but the methods, although conceptually similar, differ significantly. The estimates are well correlated at ward level (r = 0.830, 0.893 and 0.926 with respect to the proportion of adults with Entry Level literacy, numeracy and English as first language respectively), but the 2003 estimates tend to be more highly polarised than those for 2011 and have far wider Credible Intervals (CIs). The wider CIs may reflect the fact that the 2003 estimates were produced by applying model parameters to *aggregate* ward-level socio-demographic data, whereas in the present project (as described in detail below) we apply those parameters to microsimulated *individual-level* data, and only then aggregate to MSOA level the number of individuals likely to reach each literacy etc. level.

3 The local area estimates were produced in September 2011, just before the 38th Local Enterprise Partnership was announced. Northamptonshire Local Enterprise Partnership comprises seven local authorities: Corby (E07000150), Daventry (E07000151), East Northamptonshire (E07000152), Kettering (E07000153), Northampton (E07000154), South Northamptonshire (E07000155), and Wellingborough (E07000156). This was the last LEP to be included in the 16/1/2012 release of the ONS's *New LEP Local Authority Comparator Profiles* (http://www.neighbourhood.statistics.gov.uk/HTMLDocs/downloads/LEP_Profiles.xlsm)[Accessed 3/4/2012]. This website will be updated to include any further LEPs, including the recently announced Buckinghamshire Thames Valley LEP, which comprises South Buckinghamshire (E07000006), Chiltern (E07000005), Wycombe (E07000007) and Aylesbury Vale (E07000004) (http://www.bis.gov.uk/policies/economic-development/leps/statistics) [Accessed 2/4/2012], Estimates for these new partnership areas (and any further additions to list) can be approximated by aggregating LA-level estimates as described in paragraph 84 on page 72.

4 All models use data LSOA-level 4th Quarter 2010 data on the number of people claiming Incapacity Benefit and Severe Disablement Allowance. This is extracted from the *Work and Pensions Longitudinal Study* (WPLS) dataset available via NOMIS (http://www.nomisweb.co.uk/). [Accessed 12/1/2012.]

is conceptually similar to, though a clear methodological advance on, the approach used to produce literacy, numeracy and ICT skills estimates using outcome data from the *2003 Skills for Life Survey*.

4. This *Technical Report* aims to provide a full description of the methods and sources used in the study. Thus following a general introduction to the approach adopted (Section 0), this report details the principal methodological and evidential issues that have arisen (Sections 0-0) and presents key findings which throw light on the insights that can be gained through small area estimation (Sections 0) . The report concludes with a guide to how the local area estimates are presented in a series of Excel data files (Section 0), and a brief glossary of key terms (Section 0).

Small Area Estimation: An Overview

5. Small area estimation aims to overcome the problem that whilst many surveys are designed and undertaken at a national level, practitioners often require information about more local areas.[5] Unfortunately, the sample size achieved by national surveys is usually far too small for direct estimation at the sub-regional level. This is certainly the case with respect to the *2011 Skills for Life Survey*. This garnered evidence for a total of 7,230 respondents, and whilst whether or not English was the respondent's first language is recorded for all respondents, only 5,824 and 5,823 respondents respectively were assigned literacy and numeracy scores. ICT competency scores for Word Processing, Spreadsheets, Email and an ICT multiple choice test were available for just 2,253, 2,228, 2,247 and 2,274 respondents respectively. To put these figures in context, estimates are required for 7,972 statistical wards, 7,932 ST wards, 7,618 Council wards and 6,781 MSOAs. Here it is not just that sample size will be far too small for direct estimation at the sub-regional level, but that the vast majority of wards and MSOAs are not even represented in the national survey.

6. Faced with such a scenario, researchers have sought to generate local estimates from national surveys using either a form of indirect standardisation or multilevel model-based approaches.[6] Indirect standardisation, although computationally straightforward, is problematic because, by applying national or sub-national prevalence rates to local populations, it assumes spatial invariance. In other words, whilst it captures 'compositional effects' – how the overall prevalence of, say, numeracy skills, may vary from place to place simply because the socio-economic composition of populations varies – it cannot capture any 'contextual' effects that may affect its local prevalence.

7. As it seems intuitively plausible that there will be some sort of contextual (area) effect on local rates of literacy, numeracy and ICT skill levels over and above those which can be predicted on the basis of a knowledge of the socio-demographic composition of populations, it is necessary to turn to multilevel model-based approaches. Such approaches interrogate survey data in order to derive models which best describe how a dependent variable (for instance numeracy skills) responds to individual ***and*** area-level predictor variables drawn from the survey and other sources. Local area estimates for the dependent variable are then calculated by applying the model's parameter estimates to the corresponding covariate values for the local areas. In effect, the goal is to 'pool' evidence from across the wider sample in order to 'enhance' local estimates.[7]

8. An early example of this approach is provided by Twigg, Moon and Jones' work estimating the prevalence of smoking at ward-level on the basis of data drawn from the *Health Survey*

---

[5] Fay, R.E. and Herriot, R.A. (1979) "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data" *Journal of the American Statistical Association*, Vol. 74, pp269-277; Dempster, A.P and Raghunathan, T.E. (1985) "Using a Covariate for Small Area Estimation: A Common Sense Bayesian Approach" in *Small Area Statistics: An International Symposium*, eds. Platek et al., New York, Wiley, pp77-90; Rao, J.N.K. (2003) *Small Area Estimation* New York, Wiley.

[6] See glossary entries for ***Indirect Standardisation*** and ***Multilevel Models***.

[7] Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, CUP.

for England (*HSfE*).[8] Their initial step was thus to use *HSfE* data to devise a model of individual smoking behaviour using both individual-level variables derived from the *HSfE* itself (such as sex, age and marital status) and area-level variables drawn from other sources linked to HSfE respondents through place of residence (such as the percentage of private rented households in an area or the percentage of households with two or more cars). The model parameters (and their interactions) were used to estimate for each ward the proportion of smokers in each combination of age, sex and marital status, and these proportions were then applied to the corresponding census counts to provide an estimate of smoking prevalence. A number of other studies, including the *Demographic mapping of the 2003 Skills for Life Survey to local areas*, have adopted similar multilevel approaches.[9]

9. The use of multilevel models which combine individual- and area-level effects represents a significant advance in small area estimation, but individual-level variables must be identically defined in both the survey from which the model was derived and the census, or other source, from which area-level covariate data are drawn. With respect to the analysis of the *2011 Skills for Life Survey*, this means that only variables which match, or can be made to match, variables available in the *2001 Census* and/or the DWP's *Work and Pensions Longitudinal Study* dataset can be used in the final model. This does constrain variable selection in the development of models for Small Area Estimation. For instance, whilst whether or not an individual speaks English as a first language would be an obvious predictor variable in a model of adult literacy skills, the fact that there is no independent and reliable information on the number of people in local areas speaking English as a first language means that this variable, although present in the survey, cannot be used for the purposes of estimating adult literacy skill levels in local areas. Fortunately, however, it is now standard practice for most surveys and censuses to elicit a well-defined set of socio-demographic characteristics and many of the questions asked in the *2011 SfL Survey* do mirror those asked in the *2001 Census*. Questions regarding benefit status, meanwhile, can be related to Lower Layer Super Output Area (LSOA)-level data on the number of people claiming benefits in the 4th Quarter of 2010 as recorded in the DWP's *Work and Pensions Longitudinal Study* dataset.

10. A more significant issue is that it is not easy within the classical ('frequentist') statistical framework to quantify the precision of small area estimates without simplifying assumptions or the use of computationally-intensive bootstrapping techniques. In other words, the calculation of confidence intervals around estimates of the number or proportion of adults with various levels of literacy, numeracy or ICT skills in each local area is far from straightforward. Recent advances have, however, made feasible an alternative statistical framework which focuses on generating 'posterior' distributions of 'possible' estimates for each small area. Thus, rather than a producing a single estimate of, say, the proportion of people with Entry Level Literacy skills in a particular area – around which a theoretically-

---

[8] Twigg, L., Moon, G. and Jones, K. (2000) "Predicting small-area health related behaviour: a comparison of smoking and drinking indicators", *Social Science and Medicine*, 50: pp1109-20.

[9] Gibson, A., Bailey, T, and Fraser, D. (2004) *Demographic mapping of the 2003 Skills for Life Survey to local areas*. Technical Report for the Department for Education and Skills, December 2004; Heady, P. *et al*. (2003) *Small Area Estimation Project Report*. Model-Based Small Area Estimation Series No.2, ONS Publication; Longhurst, J., Cruddas, M. and Goldring, S. (2005) *Model-based Estimates of Income for Wards, 2001/02: Technical Report*. Published in Model-Based Small Area Estimation Series, ONS Publication; Bajekal, M., et al.. (2004) *Synthetic estimation of healthy lifestyles indicators: Stage 1 report*. National Centre for Social Research. (Available at http://old.iph.ie/files/file/Synthetic_Estimation_Stage_1_Report.pdf.) [Accessed 12/2/2012.]

derived confidence interval is placed – the goal is to define what is termed a 'posterior distribution' of many simulated possible outcomes *given the data being modelled*.

11. Advocates of this *Bayesian approach* to statistical modelling argue that it is more appropriate to focus in this way on the conditional ('posterior') distributions of unknown quantities – i.e. conditioning the model on the data – rather than, as in conventional 'frequentist' modelling, by conditioning the model on the basis of the distribution of a test statistic which assumes a range of unseen possibilities for the data. In brief, in the Bayesian approach parameters are thought of as 'random quantities' (rather than fixed constants as in classical statistics) so that the statistical model sought is not just the *likelihood function*, which is the probability density of the data $(y = y_1,...y_n)$ 'given' the parameter values $(\theta = \theta_1,...\theta_n)$, i.e. $\mathrm{P}(y \mid \theta)$, but rather the joint probability distribution for both the data and the parameters, i.e. $\mathrm{P}(y,\theta)$.

12. $\mathrm{P}(y,\theta)$ is linked to the likelihood function via $\mathrm{P}(y,\theta) = \mathrm{P}(y \mid \theta)\mathrm{P}(\theta)$, where $\mathrm{P}(\theta)$ is known as the ***prior*** probability distribution for the parameters because it expresses uncertainty about θ before taking the data into account. It is usually chosen to be 'non-informative'. Bayes' Theorem then allows the ***posterior*** probability distribution for the parameters to be derived given the observed data:

$$\mathrm{P}(\theta \mid y) = \frac{\mathrm{P}(y \mid \theta)\,\mathrm{P}(\theta)}{\mathrm{P}(y)} = \frac{\mathrm{P}(y \mid \theta)\,\mathrm{P}(\theta)}{\int_{\theta} \mathrm{P}(y \mid \theta)\,\mathrm{P}(\theta)\,d\theta}$$

13. In other words, the 'posterior' is proportional to 'likelihood' × 'prior' – the denominator being a normalising constant independent of the parameters. All information concerning the parameters (and functions of them, such as predictions) can then be derived for the relevant posterior distribution. Whilst in principle this offers a very general and flexible approach to statistical modelling which is capable of handling very complex modelling frameworks, the problem has always lain with the complex integrations required to evaluate the denominator involved in the expression for the posterior distribution. Indeed, whereas many modern 'real-world' implementations are faced with large numbers of random effects (as in the present instance where we have one random effect for each of the 1,516 MSOAs covered by the survey), in practice all but the very simplest integrations are, to all intents and purposes, mathematically intractable.

14. There are many numerical methods available for integration (for instance the GLLAMM package for STATA uses adaptive quadrature), but the 'curse of dimensionality' makes such approaches highly problematic for many 'real-world' applications. For instance, given t random effects and p evaluations (e.g. Quadrature points), one needs p to the power of t evaluations to fit a model – a clearly impossible target given the number of random effects met in most Small Area Estimation implementations. Over the past decade, however, the development of Markov chain Monte Carlo (McMC) simulation techniques – and computers with sufficient power to carry out the necessarily intensive calculations – has overcome the

need for complex numerical integration and made Bayesian modelling of complex situations involving many parameters a practical feasibility.[10]

15. At the heart of the McMC simulation-based approach is the construction of Markov chains with particular conditional probability density functions (i.e. the probability of one parameter given all the other parameters at their currently estimated values and the data) as their equilibrium distribution. We draw a random sample from the conditional probability density function and this draw becomes the new value of that parameter and the simulation continues to iterate. The McMC simulation is run for a long time so that it converges and sample values are collected. If such samples are numerous and the chain has properly converged then they provide virtually complete information about the required posterior distribution. A simple analogy here is that one could, for example, take a conventional regression model, with slope, intercept and variance/covariance matrix, and simulate parameter values. Important calculations (such as the prediction interval) could be derived from these simulations. McMC gives us these values directly.

16. The principal difficulty with McMC is that, in addition to ensuring the quality of model fit, we have to ensure that the algorithm has behaved correctly. Good practice suggests fitting the model several times using different starting points to ensure that the simulation converges to the same values (these models can be quite complex and we need to be sure the algorithm doesn't find a local "best" solution at the expense of a global "best" solution). Moreover, in order to avoid potential autocorrelation in the simulated values, it is also necessary to 'thin' the converged simulations by throwing away nine values in ten so that what is left is an independent random sample from the distribution of interest. With multiple McMC runs, we can then compare the different sets of simulated values to check that the simulation has reached a viable settled state.

17. McMC algorithms (such as the Gibbs sampler) have made Bayesian modelling of complex situations involving many parameters a practical feasibility.[11] The small area estimation problem provides just such a situation, and the Bayesian approach, combined with associated McMC techniques, provides a unified and flexible framework within which suitable multilevel models (involving both individual and area level covariates and both fixed and random effects) can be fitted to individual survey data and then used to generate posterior predictive distributions for small area estimates.

Bayesian Mixed-Effects (Multilevel) Models for Survey Data

18. The generality and flexibility of the Bayesian approach means that it can cope with a wide range of problems including, as in the present instance, the specification of mixed-effects or multilevel models. Multilevel models are so called because of the hierarchical (or multilevel) structure by which the data have been collected and/or within which processes are presumed to operate. Individuals are thus 'nested' within areas and their literacy, numeracy and ICT skills are assumed to be a function of both their individual social-demographic characteristics and aspects of the group of which they are a part (in this instance, their MSOA population).

---

[10] Withers, S.D. (2002), "Quantitative methods: Bayesian inference, Bayesian thinking", *Progress in Human Geography*, 26(4): pp553-66.

[11] Congdon, P. (2001), *Bayesian Statistical Modelling*. Chichester: Wiley; Congdon, P. (2003), *Applied Bayesian Modelling*. Chichester: Wiley.

19. Multilevel models are of particular importance in small area estimation because the number of individuals in each area is too small to allow for valid model parameter estimation without the ability to "borrow strength" from individuals in other areas. If this is done without accounting for correlation induced by similar area characteristics, independence assumptions (formally 'exchangeability assumptions' in a Bayesian setting) are violated. Multilevel models with 'fixed' (individual-level) and 'random' (area-level) effects – hence 'mixed-effects models' – have thus become increasingly popular in the last two decades. Precisely because the approach takes into account all uncertainty associated with unknown model parameters, the advantages of adopting a fully Bayesian approach to small area estimation have begun to be aired in the literature. Moura and Migon provide a relatively recent summary.[12]

20. The key point is not that multilevel models cannot be fitted any other way (some can), but that the Bayesian approach is more straightforward, can deal with a wider range of models and provides more comprehensive information about the estimates generated. That is to say, by adopting a Bayesian approach and thereby modelling the full posterior predictive distribution of estimates (in effect generating a large number of independent estimates of, for instance, the number of adults with Level 2 and above literacy skills) it is possible to derive empirically both a 'point estimate' of the number of adults with Level 2 and above literacy skills (the mean of the posteriors for a given MSOA) and a 95% 'credible interval' around that point estimate (the range within which 95% of the posterior estimates lie). This literally defines the range within which we are 95% certain the true value lie.

21. We have thus adopted a Bayesian approach to the calculation of the number and proportion of adults in local areas with various levels of literacy, numeracy and ICT skills, as well as the number and proportion of people who do not speak English as a first language (whom, for convenience, we will henceforth term ESOL (English Spoken as an Other Language) adults). As described in the next sections, we have modelled these phenomena on the basis of data drawn from the *2011 Skills for Life Survey*, and applied the derived parameter estimates and their distributions to corresponding covariate values for local areas (as drawn from the *2001 Census* and the DWP's *Work and Pensions Longitudinal Study* dataset). To accomplish this we have used the public domain and widely used *rjags* software[13] – a program for Bayesian analysis of complex statistical models using Markov chain Monte Carlo (McMC) techniques – with data pre-processing and post-processing being carried out using the R statistical software package[14] and MySQL[15].

---

[12] Moura, F.A.S. and Migon, H.S. (2002). "Bayesian spatial models for small area estimation of proportions", *Statistical Modelling*, 2: pp183-201.

[13] Martyn Plummer (2011). *rjags: Bayesian graphical models using MCMC. R package version 3-3*. (Available at http://CRAN.R-project.org/package=rjags.) [Accessed on 12/2/2012.]

[14] R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. (Also see http://www.R-project.org.) [Accessed on 12/2/2012.]

[15] David A. James and Saikat DebRoy (2010). *RMySQL: R interface to the MySQL database. R package version 0.7-5.*(Available at http://CRAN.R-project.org/package=RMySQL.) [Accessed on 12/1/2012.]

Skills for Life Small Area Estimation: Methodological Considerations

22. Our approach to Small Area Estimation is therefore based on the specification of Bayesian hierarchical models using data drawn from the *2011 Skills for Life Survey* and the application of model parameter distributions to corresponding covariate values for local areas. The dependent variables to be modelled are, first, whether or not individuals speak English as a first language, for which we use a ***multilevel logistic regression model***, and second, a series of skills levels for literacy, numeracy and the four ICT skill domains, for which we use ***multilevel ordinal logistic regression models***.[16] The posterior distributions themselves were obtained using the standard McMC practice of allowing and discarding a 'burn-in' of 10,000 iterations, with the following 10,000 iterations being thinned by a factor of 10 to return a posterior distribution of 1,000 estimates for each non-reference factor. Visual checks of trace plots, along with the formal application of Gelman and Rubin's diagnostic tests,[17] confirmed that all posteriors had converged to a steady state.

23. The logistic regression model uses a two category response y_{ij} where $y_{ij}=1$ if individual i within MSOA j speaks English as a first language and $y_{ij}=0$ if they do not, and where;

$$y_{ij} \sim Bernoulli\left(p_{ij}\right); \qquad p_{ij}=\frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})}$$

24. With η_{ij} specified as a linear predictor βX comprising individual-level and upper(MSOA)-level variables[18]. Specifically (see below for details of variable selection);

$$\eta_{ij}=\beta_{0j}+\beta_1\cdot sex_j+\beta_2\cdot ethnicity_j+\beta_3\cdot birthplace_j+\beta_4\cdot quals_j+ \beta_5\cdot tenure_j+\beta_6\cdot occupation_j+\beta_7\cdot benefits_j$$

$$\beta_{0j}=\gamma_0+\gamma_1\cdot lowincome_j+\varepsilon_j$$

$$\varepsilon_j=N(0,\sigma^2)$$

25. The ordinal logistic regression models used with respect to predicting literacy, numeracy and ICT skill levels, meanwhile, are specified as cumulative link models. Following standard practice,[19] we thus have an ordinal response $y_{ij} \sim Categorical(p_{ijk})$ for individuals

---

[16] We abandoned a secondary goal of modelling literacy, numeracy and ICT skill levels specifically for those who did not speak English as a first language once it became apparent that the survey contained only 610 such people. This is an insufficient number upon which to generate reasonable local estimates of the number and proportion of people in up to six skill levels across 6,781 MSOAs.

[17] Gelman, A and Rubin, DB (1992) "Inference from iterative simulation using multiple sequences", *Statistical Science*, 7, pp457-511; Brooks, SP. and Gelman, A. (1997) "General methods for monitoring convergence of iterative simulations", *Journal of Computational and Graphical Statistics*, 7, pp434-455.

[18] The variables are coded in the model using treatment contrasts (also known as indicator contrasts). The reference levels for each variable are given in Table 3 to Table 9, which give full details on the variables and parameter values for each model.

[19] McCullagh, P (1980) "Regression Models for Ordinal Data" *Journal of the Royal Statistical Society. Series B (Methodological),* Vol.42, pp109-42.

$i = 1,\ldots,n$ in MSOAs $j = 1,\ldots,m$ and with skills categories $k = 1,\ldots,K$. The ordinal response is constrained so that;

$$\sum_{k=1}^{K} p_{ijk} = 1$$

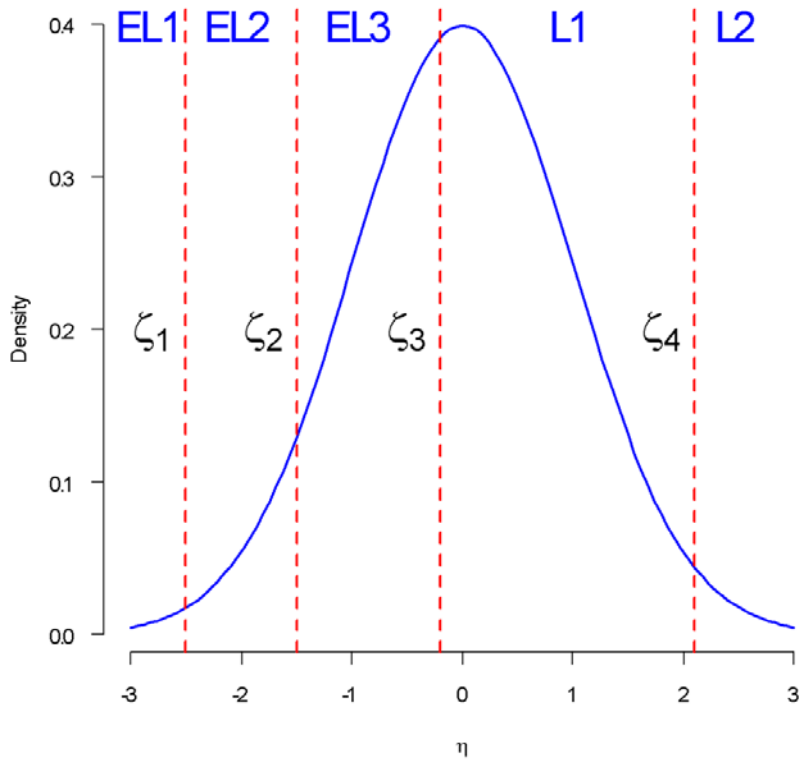
26. The key idea here is that rather than attempting to model a series of discrete factors, y_{ij} is seen as a coarsened realisation of an underlying continuous latent variable with an assumed logistic distribution.

$$p_{ijk} = \frac{\exp(\zeta_k - \eta_{ij})}{1+\exp(\zeta_k - \eta_{ij})}$$

27. We relate p_{ijk} to a linear predictor. This is done in two stages. First we assume and specify a linear predictor containing both individual-level (βX) predictors as well as an upper-level predictor $\psi[l]$. To this we then apply cut points ζ_k which define the boundaries between *K* individual skills levels in a cumulative manner, hence the alternative title "cumulative model". The upper (area-level variable) is defined as 'lowincome' which, as discussed below, is the only upper level variable included in the model. When fitting the model using McMC, therefore, we have to estimate the values of the individual- and group(MSOA)-level parameters and the cut points ζ_k.

28. We can sketch this model as in Figure 1 below. The linear predictor can be described as defining a latent variable, shown as a solid blue curve and gives the value of $\eta_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \psi[l]$ for each individual with their various characteristics described by *X*. For each individual, the cut points ζ_k (denoted in the sketch by vertical dotted red lines) act as a guide to defining the cumulative log odds. For example, the values of ζ here are -∞, -2.5, -1.5, -0.2, 2.2, +∞.

Figure 1 Illustrative Sketch of application of cut points



29. If we had a simple binomial logistic regression we would estimate the log odds for an individual as $\beta_0 + \beta_1 x_1 + \beta_2 x_2 ... + \beta_{p-1} x_{p-1} + \psi[l]$. For the proportional odds logistic regression we estimate the log odds for an individual in a cumulative way, such that the log odds of having EL1 or below skills relative to EL2, EL3, L1 or L2 and above skills is given by $\zeta_1 - \beta_1 x_1 + \beta_2 x_2 ... + \beta_{p-1} x_{p-1} + \psi[l]$, whilst the log odds of having EL2 or below relative to EL3, L1 or L2 and above skills above is given by $\zeta_2 - \beta_1 x_1 + \beta_2 x_2 ... + \beta_{p-1} x_{p-1} + \psi[l]$. There are similarities therefore between the role of β_0 in a binary logistic regression model and the sequence of ζ_k in an ordinal logistic regression model.

30. If we take as an illustration an individual whose value of η_{ij} was equal to one, we would then estimate their log odds of having EL1or below skills relative to EL2, EL3, L1 or L2 and above skills as -2.5 −1=-3.5. We would estimate their log odds of having EL2 or below skills relative to EL3, L1 or L2 and above skills as -1.5−1, i.e., -2.5 and so on. We can calculate the cumulative probabilities using the inverse logistic transformation exp(θ)/(1+exp(θ)) to indicate that the probability of having level EL1 or below skills is 0.0293. As the probability that they will have EL2 or below skills is 0.0759, the probability of having EL2 skills can be calculated as 0.0759−0.0293=0.0466.

31. These are standard models for dealing with ordinal data.[20] The key assumption is that the effect of the predictor variables is the same at different levels of the model (e.g., the effect of age or gender will be the same for the log odds of being rated EL1 or below relative to EL2 as it is for the effect of being rated EL1, EL2, EL3, L1 or L2 and above given all other levels). We assessed this "proportional odds" assumption by fitting separate binary logistic regression models for each cumulative level and checking for consistency of parameter estimates.

32. As with Binary logistic regression models, it is safe to assume that overdispersion exists. The upper level of the multilevel regression model is a model for overdispersion of which some is "explained" by modelling against MSOA level covariates and some is incorporated in the upper level random effect.

33. These models are implemented with respect to MSOAs (rather than any of the possible ward-based geographies) because it is for MSOAs that we have the most complete and up-to-date area data. In particular, use of the ONS's mid-year age-sex population estimates[21] ensures that our estimates are applied to the best available estimates at the time of analysis regarding the demographic structure of contemporary local populations. As discussed below, however, it should be noted that, excepting data on benefit recipients taken from 4th Quarter 2010 DWP returns recorded in the *Work and Pensions Longitudinal Study* dataset, all other data describing the socio-economic characteristics of MSOAs are drawn from the *2001 Census* and weighted to fit the ONS's 2009 mid-year age-sex estimates.

---

[20] Johnson, V. E. and J.H. Albert, James H. (1999) *Ordinal Data Modelling*. Springer, NY.

[21] Office for National Statistics, *Mid-2009 MSOA Quinary Estimates Revised (experimental)*. (Available at http://www.ons.gov.uk/ons/rel/sape/soa-mid-year-pop-est-engl-wales-exp/mid-2009-release/msoa-quinary-estimates.zip.) [Accessed 12/1/2012.] Although more recent estimates are now available, these were the most up-to-date statistics available to us when the analysis was undertaken.

34. The *2011 Skills for Life Survey* comprises 7,230 individuals drawn from a total of 1,516 (of 6,781) MSOAs. Individual-level effects are estimated using data drawn from the survey itself, but area-level effects use data collected and published as part of the construction of the *2010 Index of Multiple Deprivation.*[22] This provides a number of up-to-date and widely-accepted composite measures describing a wide range deprivation domains, as well as a number of individual datasets used in the construction of those measures. These can be utilised because each individual in the survey is assigned to their MSOA of residence. Table 1 below lists the individual level variables (and their factors) considered when developing the models, whilst Table 2 below lists the MSOA-level variables.

[22] McLennan, D., *et al.* (2011) *The English Indices of Deprivation 2010*, Department for Communities and Local Government. (Available at http://www.communities.gov.uk/publications/corporate/statistics/indices2010.) [Accessed 12/1/2012.]

Table 1 Individual level variables and factors available for modeling

Variable	Factors
Sex (2 factors)	Male; Female
Age (4)	16-34; 35-49; 50-59; 60-65 (although see footnote 29 below)
Household Type (4)	Single Adult; Single Parent (1 adult with 1+ children); Adult only family (2+ adults, no children); Adults with children (2+ adults with children)
Couple (2)	Not in couple; In couple
Ethnicity (6)	White; Mixed; Black/Black British; Asian/Asian British; Chinese; Other
Birth Place (2)	Not born in UK; Born in UK
Highest Qualification (6)	No qualifications; Unknown qualifications; Level 1; Level 2; Level 3; Level 4 or 5
General Health Status (5)	Very Poor; Poor; Fair; Good; Very Good
Limiting Long-term Illness (2)	No LLTI; Has LLTI
Tenure (3)	Owner Occupier; Social Rented; Privately Rented
Occupation (SOC2000) (10)	Not applicable/unable to classify; Managers & Senior Officials; Professional Occupations; Associate Professional & Technical Operations; Admin. & Secretarial Occupations; Skilled Trades Occupations; Personal Service Occupations; Sales & Customer Service Occupations; Process, Plant, & Machine Operatives; Elementary Occupations
Economic Activity Status (4)	Employed; Unemployed; Student (Econ Active & Non-Econ Active); Econ Inactive
Benefit Status: JSA (2)	No Job Seekers Allowance; Receives Job Seekers Allowance
Benefit Status: IB or SDA (2)	No Incapacity Benefit or Severe Disablement Allowance; Receives Incapacity Benefit or Severe Disablement Allowance
Benefit Status: DLA (2)	No Disability Living Allowance ; Receives Disability Living Allowance
Benefit Status: IS (2)	No Income Support; Receives Income Support
Benefit Status: HB or CTB (2)	No Housing Benefit or Council Tax Benefit ; Receives Housing Benefit or Council Tax Benefit

Table 2 MSOA-level variables available for modeling†

2010 Index of Multiple Deprivation Score	
1	Overall Index of Multiple Deprivation Score
IMD2010 Domain Scores:	
2	Income Deprivation Domain Score (2010 IMD)
3	Employment Domain Score (2010 IMD)
4	Health Deprivation & Disability Domain Score (2010 IMD)
5	Education, Skills & Training Deprivation Domain Score (2010 IMD)
6	Barriers to Housing & Services Domain Score (2010 IMD)
7	Crime Domain Score (2010 IMD)
8	Living Environment Deprivation Domain Score (2010 IMD)
Component Indicators used to construct the IMD2010	
9	Proportion population not entering higher education
10	Road distance to a supermarket or convenience store
11	Road distance to a primary school
12	Acute Morbidity (Age-sex standardised rate of emergency admission to hospital)
13	Measure of adults < 60 with mood (affective), neurotic, stress-related & somatoform disorders
14	Proportion individuals deemed to be income deprived
15	Proportion individuals deemed employment deprived

† Detailed descriptions of these variables and their construction are available in McLennan, D., et al. (2011) *The English Indices of Deprivation 2010*, DCLG.

Variable Selection and Model Fitting

35. Not all available variables were utilised in the models. Parameter selection (including a systematic search for possible interaction effects) was undertaken using standard automatic selection procedures, including via the lasso[23] and stepwise selection based on minimising the Akaike Information Criterion (AIC). This yielded a list of candidate predictor variables. A manual process was then followed to obtain a minimal set of potential variables. This particularly affected the ICT skills models in that these differed only in terms of the automatically chosen benefit indicator. As the use of a different benefit indicator for each model would have incurred a very significant additional computational overhead

[23] Tibshirani, R. (1996) " Regression shrinkage and selection via the lasso", *J. Royal. Statist. Soc B.*, Vol. 58, No. 1, pp267-288.

during the microsimulation stage, we compared the model fit for each outcome variable with each possible benefit indicator, and chose one common variable that was the best compromise.

36. The resulting candidate models were then fitted using approximate Bayesian methods to ensure that each model was suitable (using the arm library of R).[24] These methods capture fixed effects only, but can be used to check that residual assumptions are valid and to compare model predicted response with the actual response. At this stage we also considered the range of potential MSOA-level predictor variables by manually fitting models with each possible predictor and sensible combination of predictors in turn. On this basis 'lowincome' (i.e. the proportion individuals in each MSOA deemed to be income deprived according to the authors of the English Indices of Deprivation 2010[25]) was identified as being a sufficient proxy for underlying differences between MSOA populations.

37. Having obtained a parsimonious set of candidate models, McMC using *rjags* was used to obtain simulated values for the posterior distributions of all parameters, given the data. As discussed below, it was upon these posterior values that we predicted responses for each person type simulated in the UK. These individuals were then apportioned to the various geographies for which estimates were required.

38. Table 3 to Table 9 below list the factors included in each of the models, along with each factor's posterior mean and Standard Error. These tables, which are also presented graphically for interpretative purposes (Figure 2 to Figure 8), illustrate the relative importance of each parameter. A large positive posterior mean attached to a particular factor, for instance being a student in any of the adult skills models, indicates that those individuals are likely to have a higher skill level relative to the reference group (in this case people in employment). Conversely, a large negative posterior mean attached to a particular factor, for instance people whose occupation is unknown or not applicable (a group almost exclusively comprising those who have never had a job which, among the young, may be because they have been unemployed since leaving school but, more usually, is presumably because they have remained at home raising a family), is indicative of a lower skill level relative to the reference group (which in this case is managers and senior officials).

39. Each factor's posterior Standard Error meanwhile, is indicative of the level of model uncertainty around that factor's posterior mean. Where the number of people in the survey is small as, for instance, with respect to those with Chinese ethnicity (n=20), there is simply too little information to allow for a precise estimate of the corresponding posterior mean. Relatively large Standard Errors can also arise if there is significant diversity of outcome relative to a given factor. Such may explain, for instance, the relatively large size of the Standard Error attached to people for whom occupation is not known or not applicable. Such people have generally lower skill levels than the reference group (managers and

---

[24] Andrew Gelman et al. (2011) *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. *R package version 1.4-14* (http://cran.r-project.org/web/packages/arm/) [Accessed 9/1/2012].

[25] McLennan, D., *et al.* (2011) *The English Indices of Deprivation 2010*, Department for Communities and Local Government, p19. (Available at http://www.communities.gov.uk/publications/corporate/statistics/indices2010technicalreport.) [Accessed 12/1/2012.]

senior officials), but there is considerable uncertainty about the relative size of that effect because of the diversity of skills possessed by people in that group.

40. It is worth noting that, by and large, the individual-level factors have the effects that might have been expected. For instance, an individual's 'highest qualification' is strongly associated with all skill sets and is included in all models except literacy. In this case, whilst highest qualification remains highly correlated with adult literacy, it does not form part of the final model because an alternative set of predictor variables offers a better fit with the data. It is worth noting here that our approach has been data-, rather than theory-, led in that we have used whatever set of predictor values minimised the AIC. We are not, in other words, seeking to develop interpretative models but rather derive the best models for predictive purposes.

41. Occupation is also strongly associated with adult skills, with only 'professional occupations' having clearly higher literacy and numeracy skills than the reference group (managers and senior officials). As might be expected, when one turns to consider ICT skills, it is the 'administrative and secretarial occupation' category which emerges with distinctively higher email and spreadsheet skills than other occupation groups. In terms of both general patterns and such specific comparisons, all the skills models have very strong face validity.

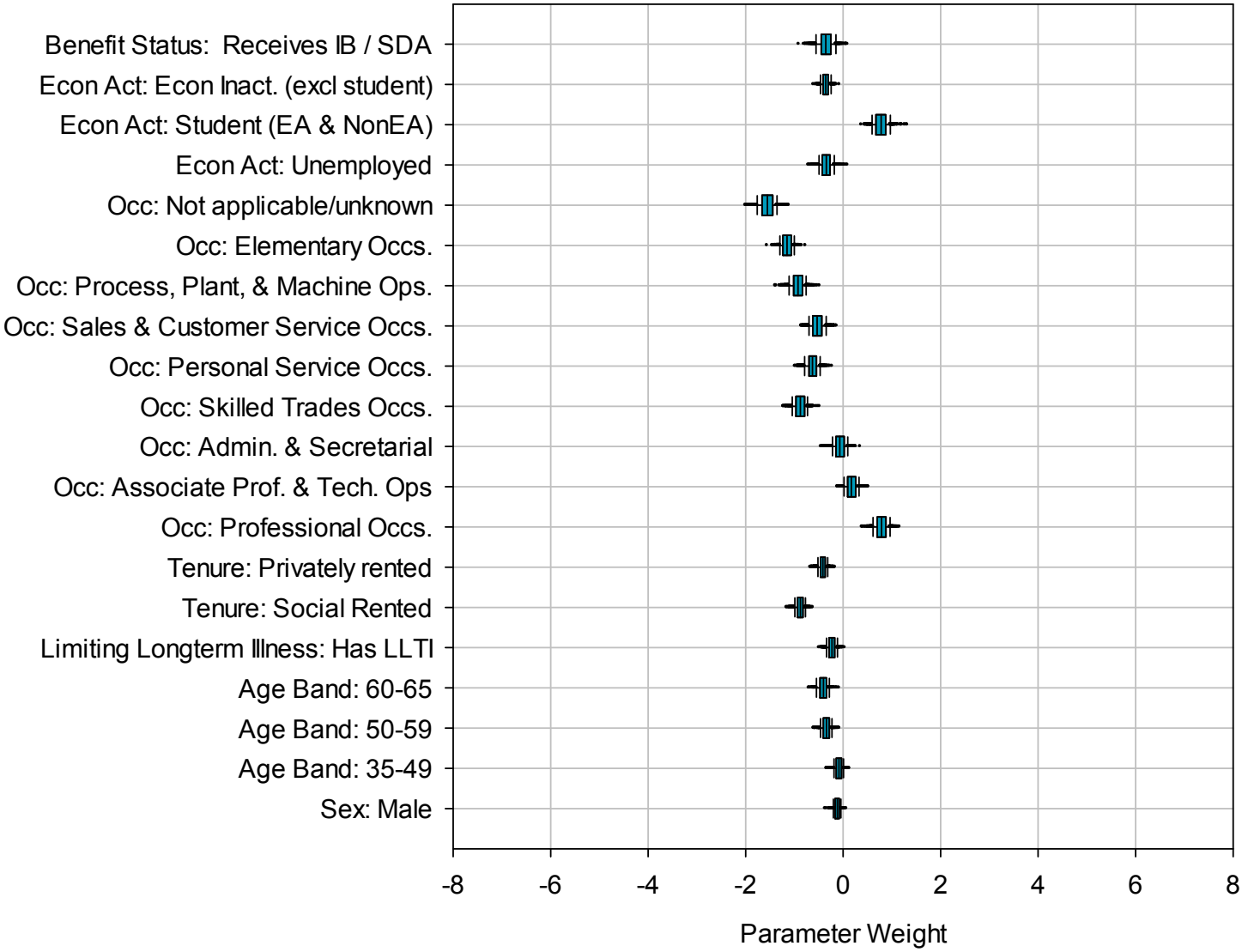
42. The English Spoken as an Other Language (ESOL) model also possesses good face validity, with non-white ethnicity being, as one would expect, very strongly associated with ESOL status (note the size of the parameter estimates on Table 9 / Figure 8). Conversely, being born in the UK is, as one might expect, strongly predictive of speaking English as a first language. Other aspects of the ESOL model are interesting without necessarily being counter-intuitive. ESOL speakers tend not to be on benefits; are more likely to live in privately-rented accommodation than in either social housing or as owner occupiers; and are least likely to be in the 'managers and senior officials' occupation category. As noted above, the parameter estimates in the linear predictor (β) represent the log-odds of an individual having the higher level of literacy relative to all lower levels. So for example, a posterior mean of -0.3521 for "Receives IB/SDA" indicates that the log odds are reduced for each category (EL2 versus EL1, EL3 versus EL1 and EL2 and so on) by -0.3521 for people in receipt of IB/SDA (against the reference group "Does not receive IB/SDA").

Table 3 Individual-level Parameter Estimates: Literacy

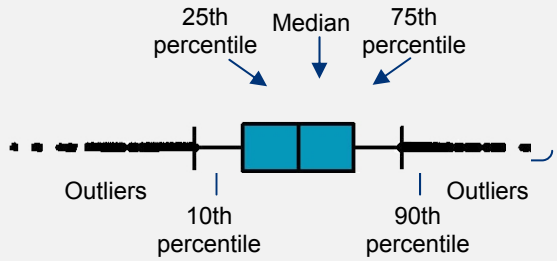
Factor [Reference Group]	Posterior Mean	Posterior Standard Error
Benefit Status:		
Receives IB / SDA [Does not receive IB / SDA]	-0.3521	0.1555
Economic Activity:		
Econ Inactive (except student) [Employed]	-0.3539	0.0799
Student (EA & NonEA) [Employed]	0.7813	0.1472
Unemployed [Employed]	-0.3413	0.1209
Occupation:		
Not applicable/unknown [Managers & Senior Officials]	-1.5567	0.1586
Elementary Occupations [Managers & Senior Officials]	-1.1489	0.1141
Process, Plant, & Machine Operatives [Managers & Senior Officials]	-0.9249	0.1351
Sales & Customer Service Occupations [Managers & Senior Officials]	-0.5257	0.1359
Personal Service Occupations [Managers & Senior Officials]	-0.6253	0.1220
Skilled Trades occupations [Managers & Senior Officials]	-0.8783	0.1223
Administrative & Secretarial [Managers & Senior Officials]	-0.0620	0.1244
Associate Prof. & Tech. Operations [Managers & Senior Officials]	0.1770	0.1186
Professional Occupations [Managers & Senior Officials]	0.7881	0.1314
Tenure:		
Privately rented [Owner Occupier]	-0.4148	0.0794
Social Rented [Owner Occupier]	-0.8767	0.0838
Limiting Longterm Illness:		
Has LLTI [No LLTI]	-0.2284	0.0864
Age Band:		
60-65 [16-34]	-0.4082	0.1002
50-59 [16-34]	-0.3431	0.0914
35-49 [16-34]	-0.0852	0.0760
Sex:		
Male [Female]	-0.1203	0.0615

Note: See paragraphs 38 and 39 for a brief account of how to 'read' this and subsequent tables. Figure 2 to Figure 8 provide a more immediate 'visual' impression of the 'significance' of the various factors included in each model.

Figure 2 Individual-level Parameter Estimates: Literacy



Note: Figure 2 to Figure 8 illustrate the nature of the posterior distributions for each factor in each model. The box plots themselves should be 'read' as follows:



Each of the figures uses the same x-axis scale (-8 to +8) to aid understanding of the relative 'significance' of the different factors. As a rule-of-thumb, where a factor's posterior distribution does not cross zero one can be reasonably certain that it is what would be conventionally regarded as statistically significant at the 5% level (2-tailed).

Table 4 Individual-level Parameter Estimates: Numeracy

Factor [Reference Group]	Posterior Mean	Posterior Standard Error
Benefit Status:		
Receives IB/SDA [Does not receive IB/SDA]	-0.3449	0.1820
Economic Activity:		
Econ Inactive (except student) [Employed]	-0.2519	0.0923
Student (EA & NonEA) [Employed]	0.6784	0.1484
Unemployed [Employed]	-0.3214	0.1444
Occupation:		
Not applicable or unknown [Managers & Senior Officials]	-0.9770	0.1797
Elementary Occupations [Managers & Senior Officials]	-0.6546	0.1229
Process, Plant, & Machine Operatives [Managers & Senior Officials]	-0.5187	0.1457
Sales & Customer Service Occupations [Managers & Senior Officials]	-0.2395	0.1443
Personal Service Occupations [Managers & Senior Officials]	-0.5701	0.1340
Skilled Trades occupations [Managers & Senior Officials]	-0.5596	0.1338
Admin. & Secretarial Occupations [Managers & Senior Officials]	0.0255	0.1206
Associate Prof.& Tech. Operations [Managers & Senior Officials]	-0.0133	0.1243
Professional Occupations [Managers & Senior Officials]	0.6472	0.1541
Tenure:		
Privately rented [Owner Occupier]	-0.3027	0.0808
Social Rented [Owner Occupier]	-0.5359	0.1028
Limiting Longterm Illness:		
Has LLTI [No LLTI]	-0.2480	0.0999
Highest Qualification:		
Level 4/5 [No Qualifications]	2.0398	0.1077
Level 3 [No Qualifications]	1.6124	0.1123
Level 2 [No Qualifications]	1.2775	0.1094
Level 1 [No Qualifications]	1.0301	0.1024
Unknown qualifications/level [No Qualifications]	0.2965	0.1691
Ethnicity:		
Other [White]	-1.6610	0.2788
Chinese [White]	-1.6046	0.6536
Asian or Asian British [White]	-1.0377	0.1814
Black or Black British [White]	-1.3438	0.1409
Mixed [White]	-0.2665	0.2837
Sex:		
Male [Female]	-0.0816	0.0713

Figure 3 Individual-level Parameter Estimates: Numeracy

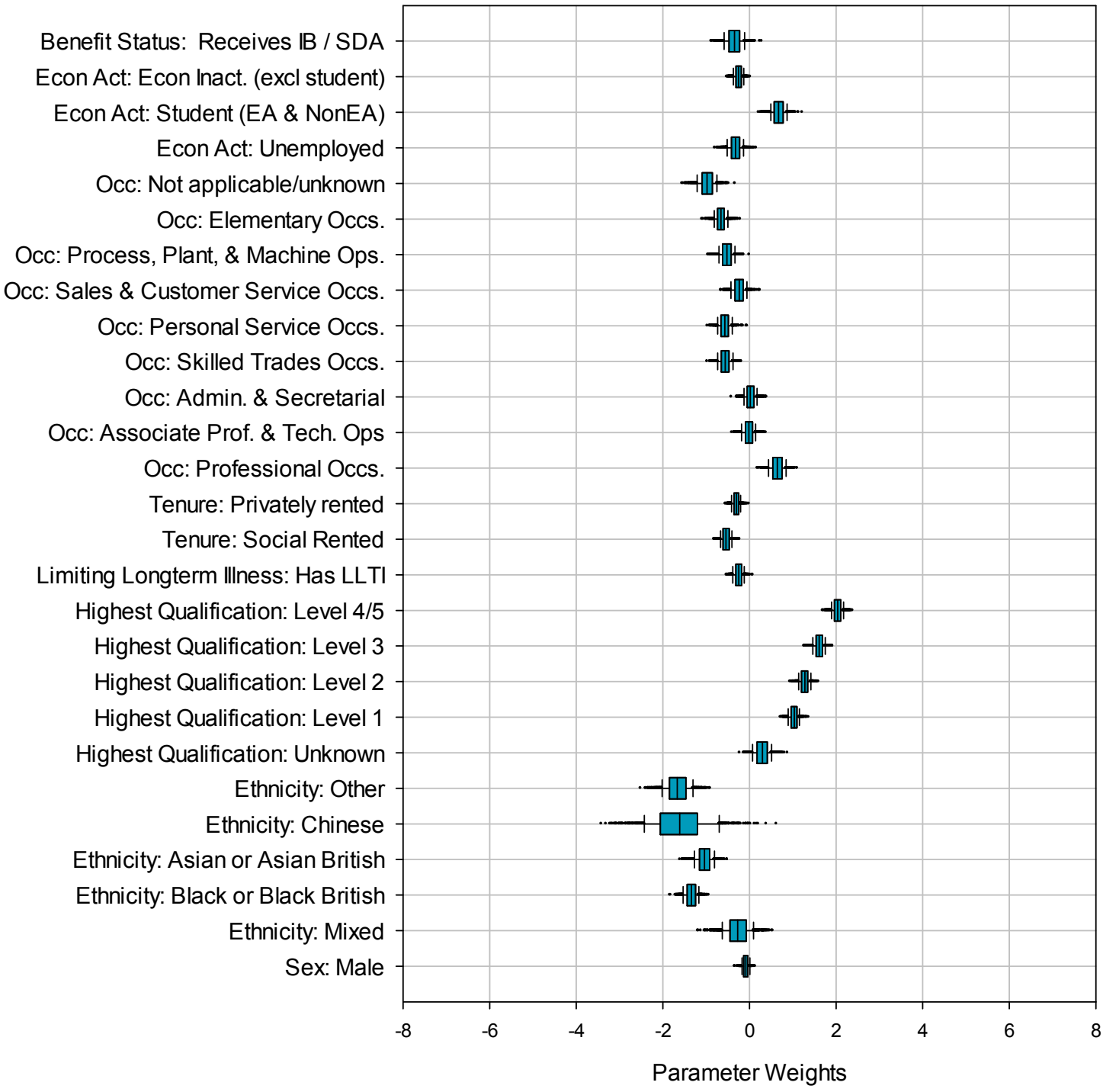


Table 5 Individual-level Parameter Estimates: ICT - Email

Factor [Reference Group]	Posterior Mean	Posterior Standard Error
Benefit Status:		
Receives IB/SDA [Does not receive IB/SDA]	0.1271	0.2894
Economic Activity:		
Econ Inactive (except student) [Employed]	-0.3687	0.1280
Student (EA & NonEA) [Employed]	0.7826	0.2340
Unemployed [Employed]	-0.2512	0.2242
Occupation:		
Not applicable or unknown [Managers & Senior Officials]	-0.4510	0.2673
Elementary Occupations [Managers & Senior Officials]	-1.2438	0.1922
Process, Plant, & Machine Operatives [Managers & Senior Officials]	-1.2114	0.2210
Sales & Customer Service Occupations [Managers & Senior Officials]	-0.8107	0.2045
Personal Service Occupations [Managers & Senior Officials]	-1.2550	0.2089
Skilled Trades occupations [Managers & Senior Officials]	-1.3549	0.1923
Administrative & Secretarial Occupations [Managers & Senior Officials]	0.5126	0.1795
Associate Prof. & Tech. Operations [Managers & Senior Officials]	-0.0182	0.1794
Professional Occupations [Managers & Senior Officials]	-0.1190	0.1866
Tenure:		
Privately rented [Owner Occupier]	-0.1786	0.1348
Tenure: Social Rented [Owner Occupier]	-0.6515	0.1463
Limiting Longterm Illness:		
Has LLTI [No LLTI]	-0.2868	0.1603
Highest Qualification:		
Level 4/5 [No Qualifications]	2.3396	0.1903
Level 3 [No Qualifications]	1.8731	0.1896
Level 2 [No Qualifications]	1.6543	0.1909
Level 1 [No Qualifications]	1.0060	0.1774
Unknown qualifications/level [No Qualifications]	0.8764	0.3144
Ethnicity:		
Other [White]	-0.1371	0.4287
Chinese [White]	-0.1440	1.4000
Asian or Asian British [White]	-0.7872	0.2753
Black or Black British [White]	-1.1891	0.2315
Mixed [White]	-0.4068	0.3735
Age Band:		
60-65 [16-34]	-1.7315	0.1894
50-59 [16-34]	-1.3704	0.1604
35-49 [16-34]	-0.7866	0.1329

Figure 4 Individual-level Parameter Estimates: ICT - Email

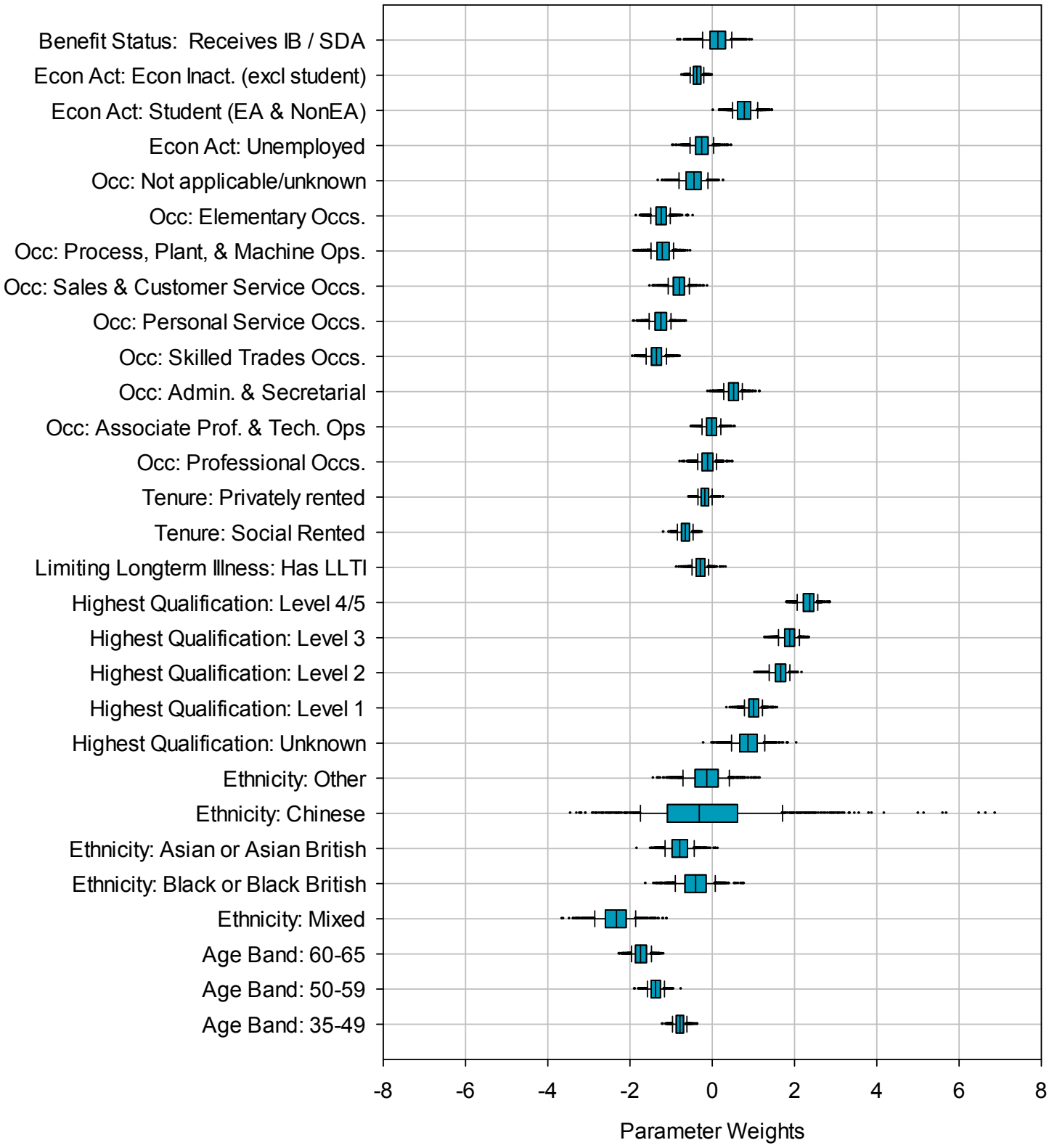


Table 6 Individual-level Parameter Estimates: ICT - Word Processing

Factor [Reference Group]	Posterior Mean	Posterior Standard Error
Benefit Status:		
Receives IB/SDA [Does not receive IB/SDA]	-0.4535	0.2283
Economic Activity:		
Econ Inactive (except student) [Employed]	-0.3162	0.1130
Economic Activity: Student (EA & NonEA) [Employed]	1.2519	0.2031
Economic Activity: Unemployed [Employed]	-0.5040	0.1843
Occupation:		
Not applicable or unknown [Managers & Senior Officials]	-0.3193	0.2410
Elementary Occupations [Managers & Senior Officials]	-1.1924	0.1659
Process, Plant, & Machine Operatives [Managers & Senior Officials]	-1.1197	0.1924
Sales & Customer Service Occupations [Managers & Senior Officials]	-0.6235	0.1680
Personal Service Occupations [Managers & Senior Officials]	-1.1942	0.1770
Skilled Trades occupations [Managers & Senior Officials]	-1.4298	0.1883
Admin. & Secretarial Occupations [Managers & Senior Officials]	0.3905	0.1611
Associate Prof. & Tech. Operations [Managers & Senior Officials]	-0.0312	0.1566
Professional Occupations [Managers & Senior Officials]	0.3432	0.1657
Tenure:		
Privately rented [Owner Occupier]	-0.0329	0.1060
Tenure: Social Rented [Owner Occupier]	-0.6481	0.1285
Highest Qualification:		
Level 4/5 [No Qualifications]	2.5124	0.1466
Level 3 [No Qualifications]	2.1510	0.1702
Level 2 [No Qualifications]	1.8021	0.1565
Level 1 [No Qualifications]	0.9942	0.1582
Unknown qualifications/level [No Qualifications]	0.6825	0.2535
Ethnicity:		
Other [White]	-1.1737	0.3830
Ethnicity: Chinese [White]	0.7930	0.9938
Ethnicity: Asian or Asian British [White]	-0.9097	0.2451
Ethnicity: Black or Black British [White]	-1.1801	0.2092
Ethnicity: Mixed [White]	-0.3353	0.3570
Age Band:		
60-65 [16-34]	-1.7705	0.1593
Age Band: 50-59 [16-34]	-1.4434	0.1278
Age Band: 35-49 [16-34]	-0.7987	0.0977
Sex:		
Male [Female]	0.2116	0.0955

Figure 5 Individual-level Parameter Estimates: ICT - Word Processing

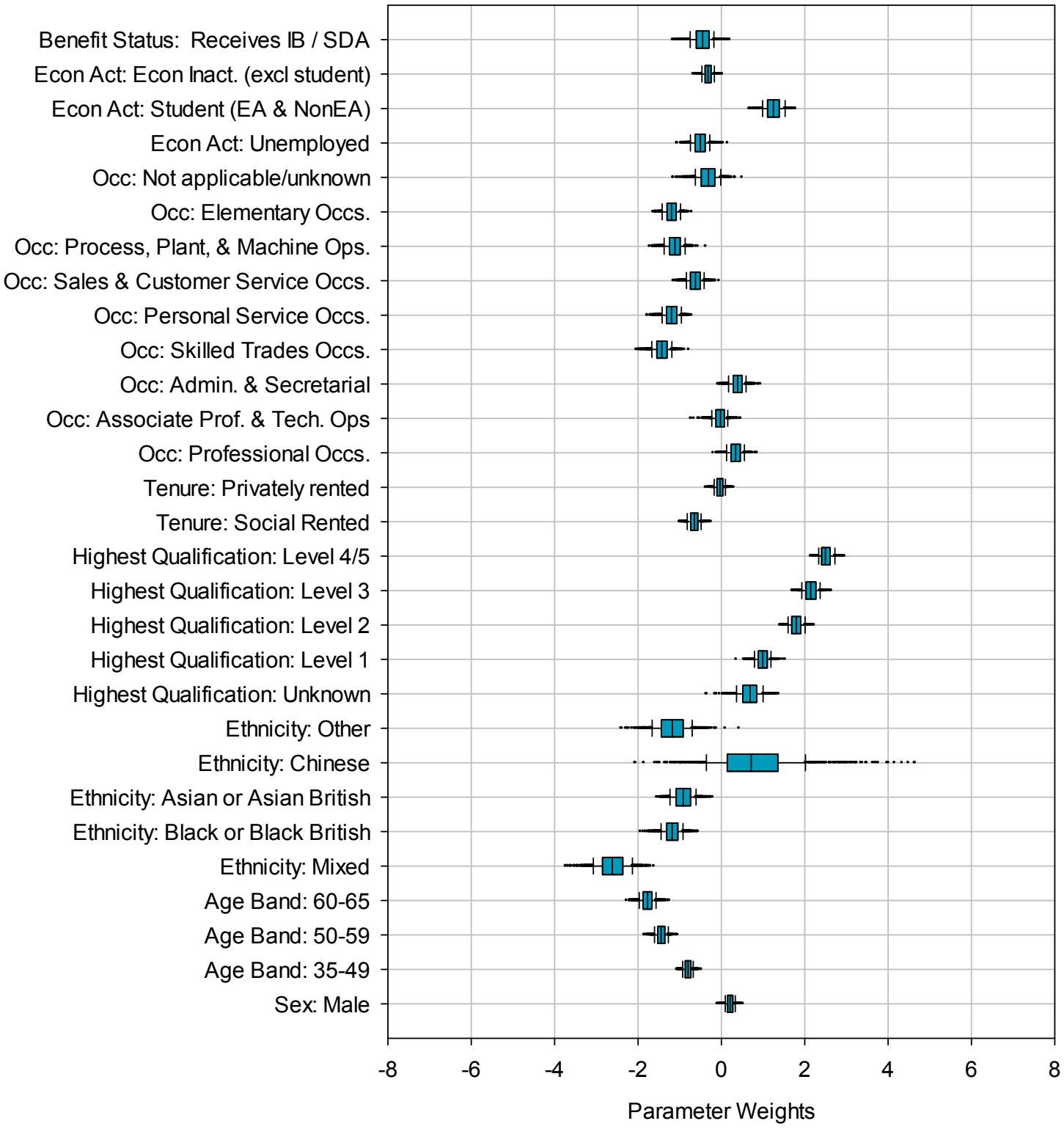


Table 7 Individual-level Parameter Estimates: ICT - Spreadsheets

Factor [Reference Group]	Posterior Mean	Posterior Stand. Err.
Benefit Status:		
Receives IB/SDA [Does not receive IB/SDA]	0.1486	0.3216
Economic Activity:		
Econ Inactive (except student) [Employed]	-0.4880	0.1349
Student (EA & NonEA) [Employed]	0.7625	0.2139
Unemployed [Employed]	-0.6451	0.2217
Occupation:		
Not applicable or unknown [Managers & Senior Officials]	-0.5619	0.2697
Elementary Occupations [Managers & Senior Officials]	-1.2631	0.1754
Process, Plant, & Machine Operatives [Managers & Senior Officials]	-1.5885	0.2390
Sales & Customer Service Occupations [Managers & Senior Officials]	-1.0206	0.2024
Personal Service Occupations [Managers & Senior Officials]	-1.6031	0.2138
Skilled Trades occupations [Managers & Senior Officials]	-1.4994	0.1936
Administrative & Secretarial Occupations [Managers & Senior Officials]	0.2933	0.1698
Associate Prof. & Tech. Operations [Managers & Senior Officials]	-0.5003	0.1475
Professional Occupations [Managers & Senior Officials]	-0.0696	0.1624
Tenure:		
Privately rented [Owner Occupier]	-0.1463	0.1226
Social Rented [Owner Occupier]	-0.6027	0.1497
Limiting Longterm Illness:		
Has LLTI [No LLTI]	-0.6124	0.1638
Highest Qualification:		
Level 4/5 [No Qualifications]	2.1599	0.1669
Level 3 [No Qualifications]	1.8548	0.1858
Level 2 [No Qualifications]	1.5639	0.1762
Level 1 [No Qualifications]	0.9283	0.1834
Unknown qualifications/level [No Qualifications]	0.5827	0.3320
Ethnicity:		
Other [White]	-0.7098	0.4158
Chinese [White]	-0.2323	1.2797
Asian or Asian British [White]	-1.3217	0.2864
Black or Black British [White]	-1.0178	0.2341
Mixed [White]	-0.2071	0.4076
Age Band:		
60-65 [16-34]	-1.8964	0.1882
50-59 [16-34]	-1.4677	0.1525
35-49 [16-34]	-0.9386	0.1221
Sex:		
Male [Female]	0.4725	0.1013

Figure 6 Individual-level Parameter Estimates: ICT – Spreadsheets

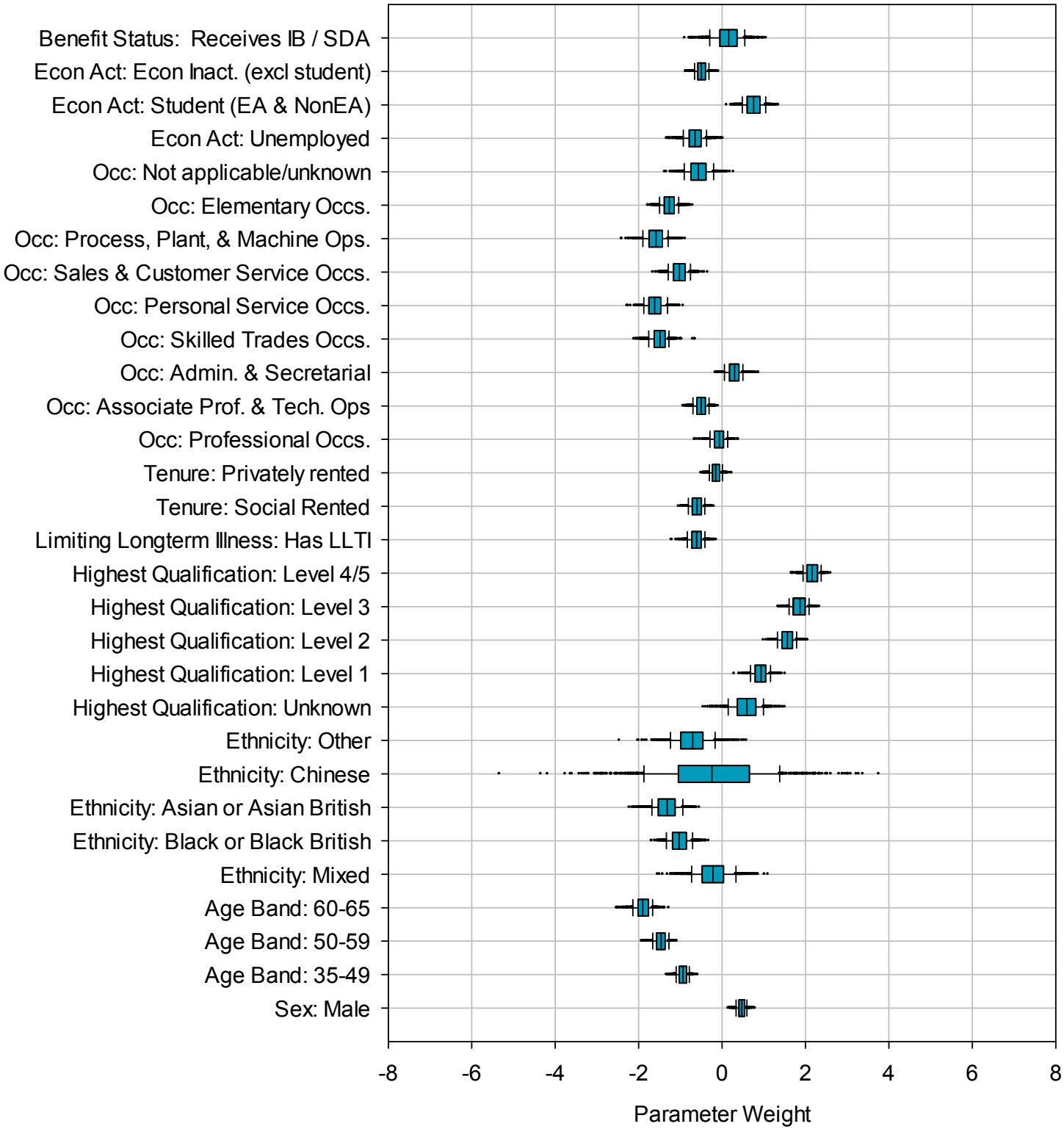


Table 8 Individual-level Parameter Estimates: ICT - Multiple Choice

Factor [Reference Group]	Posterior Mean	Posterior Stand. Err.
Benefit Status:		
Receives IB/SDA [Does not receive IB/SDA]	-0.1889	0.2302
Economic Activity:		
Econ Inactive (except student) [Employed]	-0.3971	0.1153
Student (EA & NonEA) [Employed]	0.9970	0.2316
Unemployed [Employed]	-0.3422	0.2075
Occupation:		
Not applicable or unknown [Managers & Senior Officials]	-0.6586	0.2522
Elementary Occupations [Managers & Senior Officials]	-1.1945	0.1972
Process, Plant, & Machine Operatives [Managers & Senior Officials]	-0.9082	0.2361
Sales & Customer Service Occupations [Managers & Senior Officials]	-0.5983	0.2119
Personal Service Occupations [Managers & Senior Officials]	-1.1571	0.2018
Skilled Trades occupations [Managers & Senior Officials]	-1.1395	0.2023
Administrative & Secretarial Occupations [Managers & Senior Officials]	0.3154	0.1992
Associate Prof. & Tech. Operations [Managers & Senior Officials]	0.0997	0.1887
Professional Occupations [Managers & Senior Officials]	0.4041	0.2124
Tenure:		
Privately rented [Owner Occupier]	-0.2070	0.1189
Social Rented [Owner Occupier]	-0.8075	0.1289
Highest Qualification:		
Level 4/5 [No Qualifications]	2.3620	0.1652
Level 3 [No Qualifications]	1.9187	0.1728
Level 2 [No Qualifications]	1.8034	0.1717
Level 1 [No Qualifications]	1.0659	0.1554
Unknown qualifications/level [No Qualifications]	0.6957	0.2563
Birth Place:		
UK [Not UK]	0.6616	0.1422
Ethnicity:		
Other [White]	-0.9746	0.3895
Chinese [White]	0.6622	1.6143
Asian or Asian British [White]	-0.7014	0.2581
Black or Black British [White]	-1.2780	0.2145
Mixed [White]	-0.1321	0.3625
Age Band:		
60-65 [16-34]	-1.1617	0.166
50-59 [16-34]	-0.7342	0.1359
35-49 [16-34]	-0.1985	0.1095
Sex:		
Male [Female]	0.2733	0.1016

Figure 7 Individual-level Parameter Estimates: ICT - Multiple Choice

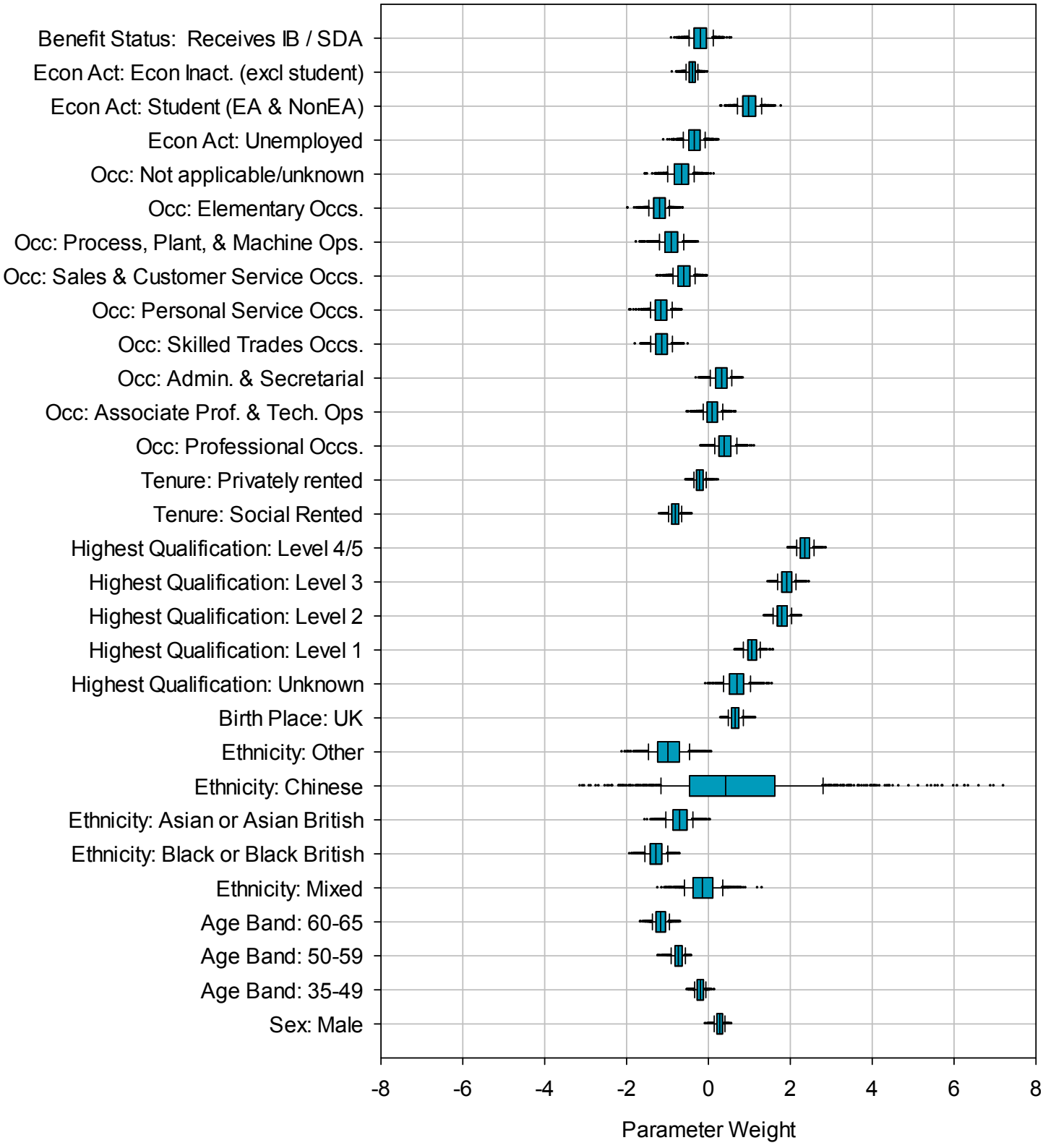
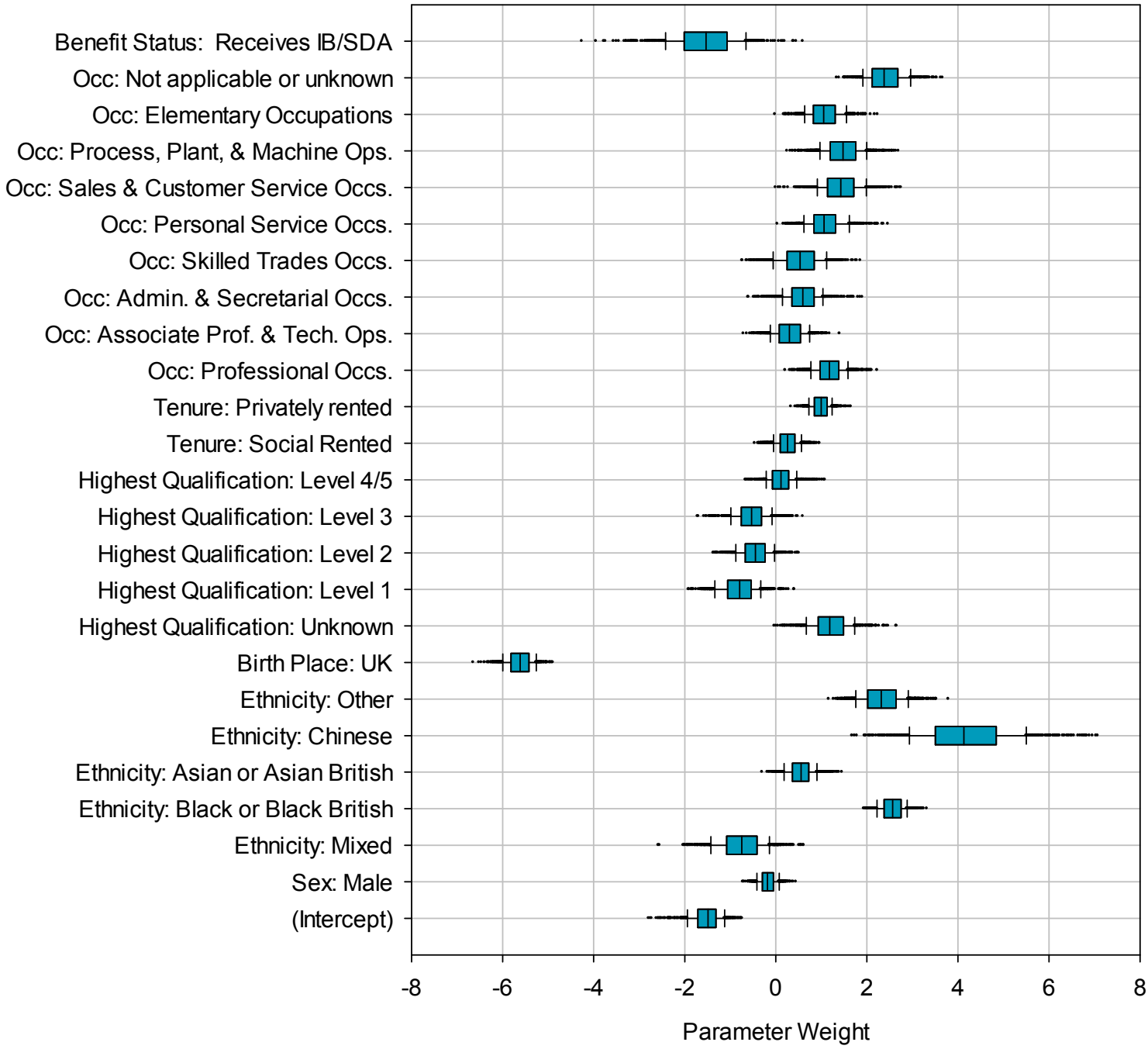


Table 9 Individual-level Parameter Estimates: English not as First Language

Factor [Reference Group]	Posterior Mean	Posterior Standard Error
Benefit Status:		
Receives IB/SDA [Does not receive IB/SDA]	-1.5491	0.6911
Occupation:		
Not applicable or unknown [Managers & Senior Officials]	2.4041	0.4046
Elementary Occupations [Managers & Senior Officials]	1.0730	0.3479
Process, Plant, & Machine Operatives [Managers & Senior Officials]	1.4774	0.4087
Sales & Customer Service Occupations [Managers & Senior Officials]	1.4344	0.4278
Personal Service Occupations [Managers & Senior Officials]	1.0898	0.3854
Skilled Trades occupations [Managers & Senior Officials]	0.5353	0.4426
Administrative & Secretarial Occupations [Managers & Senior Officials]	0.5944	0.3700
Associate Prof. & Tech. Operations [Managers & Senior Officials]	0.3047	0.3315
Professional Occupations [Managers & Senior Officials]	1.1850	0.3177
Tenure:		
Privately rented [Owner Occupier]	0.9888	0.2001
Social Rented [Owner Occupier]	0.2579	0.2406
Highest Qualification:		
Level 4/5 [No Qualifications]	0.1234	0.2733
Level 3 [No Qualifications]	-0.5365	0.3529
Level 2 [No Qualifications]	-0.4537	0.3323
Level 1 [No Qualifications]	-0.8089	0.3877
Unknown qualifications/level [No Qualifications]	1.2054	0.4287
Birth Place:		
UK [Not UK]	-5.6237	0.2852
Ethnicity:		
Other [White]	2.3305	0.4424
Chinese [White]	4.1907	0.9878
Asian or Asian British [White]	0.5527	0.2725
Black or Black British [White]	2.5610	0.2523
Mixed [White]	-0.7577	0.4960
Sex:		
Male [Female]	-0.1729	0.1857
(***Intercept)***	-1.5224	0.3305

Figure 8 Individual-level Parameter Estimates: English not as First Language



43. In addition to the individual 'fixed effects', each multilevel model includes an upper-level regression model component (as described in paragraph 24 above). This aims to capture variations in skill levels that are due to *contextual* as opposed to *compositional* effects – in other words, the extent to which the characteristics of places (or, more specifically, of MSOA populations) can explain variations in skill levels (or ESOL status) over and above that which can be explained with reference to an individual's socio-demographic characteristics. Given the desire for a parsimonious set of models, parameter selection determined that, of the upper-level variables listed in Table 2 above, it was appropriate and sufficient to include only the 'low income' variable. This, on its own, stands as a proxy for underlying differences between MSOA populations and very little additional model power (i.e. reduction in AIC) could be achieved through the inclusion of additional MSOA-level variables. The posterior means and Standard Errors for the 'low income' variable as fitted in each of the models is given in Table 10 below, indicating that whilst some sort of geographically-defined low-income effect can be identified for most of the models, it is never what would traditionally be considered statistically significant at the 95% level.

Table 10 Upper-level Parameter Estimates (Low Income): All Models

Model	Posterior Mean	Posterior Standard Error
Literacy	0.3721	0.2639
Numeracy	0.4108	0.3164
ICT-Email	-0.1619	0.3329
ICT - Word Processing	-0.3641	0.4196
ICT - Spreadsheets	-0.3920	0.3469
ICT - Multiple Choice	0.0655	0.3396
ESOL	0.5505	0.8343

44. The final statistics to be modelled are the 'cut points' used to divide the underlying continuous latent variable into discrete skills levels.[26] Presented in Table 11 (and illustrated by Figure 9), the cut points are well defined with respect to literacy, numeracy, word processing and spreadsheet skills. With respect to the multiple choice test on ICT skills and email skills, the models offer much less definition, at least with respect to the boundaries between the lower skill levels. This means we are necessarily less certain of how the number and proportion of adults with lower

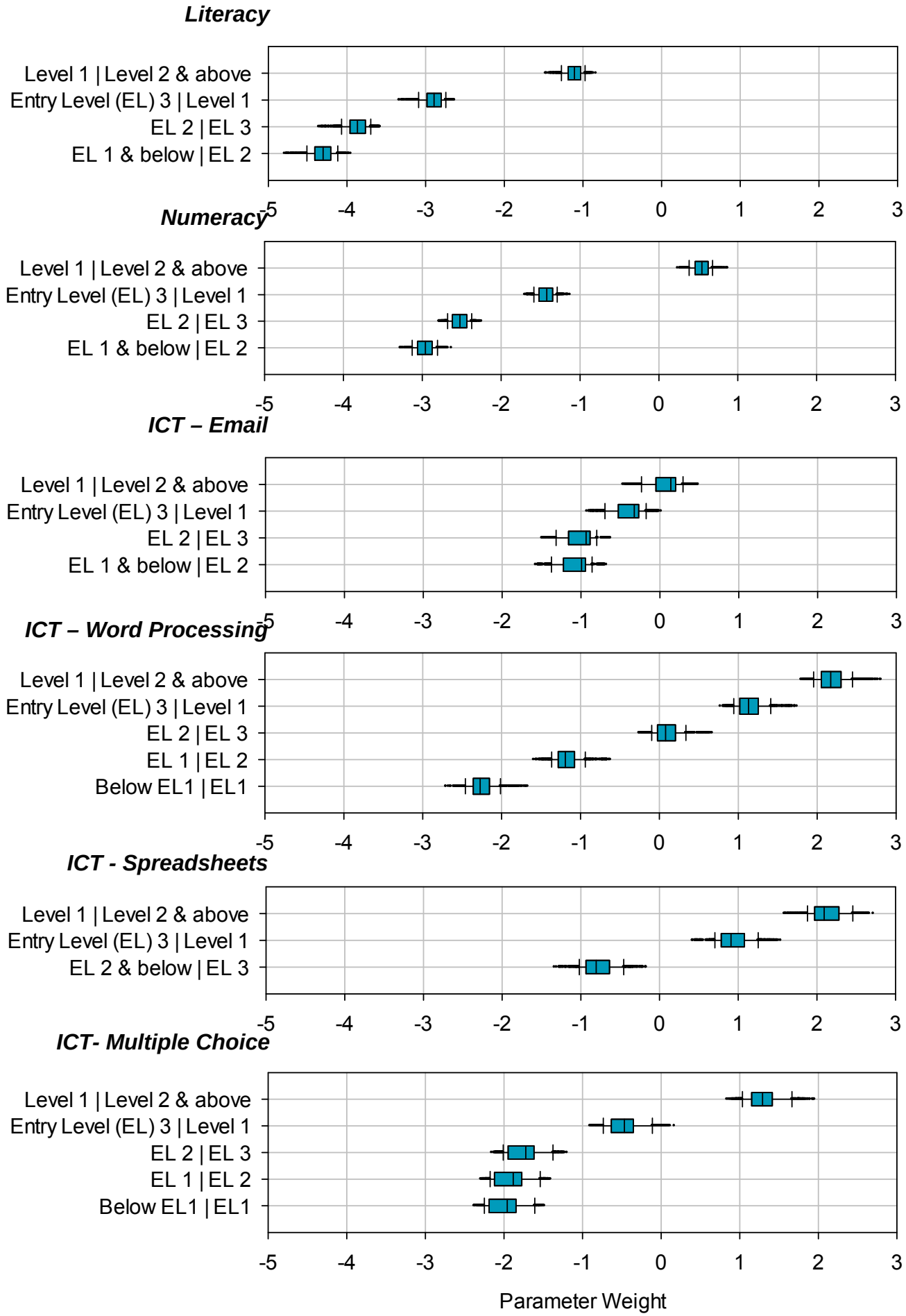
[26] Note that these 'cut points' do not refer to the assessment thresholds used to categorise, say, survey-based literacy scores into skill levels, but rather refer to the thresholds we use to subdivide into discrete categories the underlying latent variable we presume relates individual person characteristics to basic skills levels. The former was imposed as part of the survey, the latter is a modelling construct as described in paragraph 27 above.

ICT and email skills splits between specific Entry Level categories, which in turn results in relatively wider CIs for these skill estimates.

Table 11 Cut Point Estimates: All 'Skill Level' models

Literacy	Posterior Mean	Posterior Standard Error
Level 1 | Level 2 & above	-1.1146	0.1169
Entry Level 3 | Level 1	-2.9019	0.1323
Entry Level 2 | Entry Level 3	-3.8756	0.1460
Entry Level 1 & below | Entry Level 2	-4.3110	0.1544
Numeracy	**Posterior Mean**	**Posterior Standard Error**
Level 1 | Level 2 & above	0.7607	0.1130
Entry Level 3 | Level 1	0.0700	0.1115
Entry Level 2 | Entry Level 3	-0.6244	0.1150
Entry Level 1 & below | Entry Level 2	-1.4351	0.1200
ICT – Email	**Posterior Mean**	**Posterior Standard Error**
Level 1 | Level 2 & above	0.0790	0.2033
Entry 3 | Level 1	-0.3838	0.2036
Entry 2 | Entry 3	-0.9956	0.2041
Entry Level 1 & below | Entry 2	-1.0549	0.2048
ICT - Word Processing	**Posterior Mean**	**Posterior Standard Error**
Level 1 | Level 2 & above	2.1898	0.1916
Entry Level 3 | Level 1	1.1554	0.1863
Entry Level 2 | Entry Level 3	0.1043	0.1802
Entry Level 1 | Entry Level 2	-1.1662	0.1758
Below Entry Level 1 | Entry Level 1	-2.2463	0.1765
ICT - Spreadsheets	**Posterior Mean**	**Posterior Standard Error**
Level 1 | Level 2 & above	2.1301	0.2207
Entry Level 3 | Level 1	0.9431	0.2192
Entry Level 2 & below | Entry Level 3	-0.7737	0.2233
ICT- Multiple Choice	**Posterior Mean**	**Posterior Standard Error**
Level 1 | Level 2 & above	1.3074	0.2208
Entry Level 3 | Level 1	-0.4624	0.2200
Entry Level 2 | Entry Level 3	-1.7413	0.2240
Entry Level 1 | Entry Level 2	-1.9061	0.2246
Below Entry Level | Entry Level 1	-1.9809	0.2259

Figure 9 Cut Point Estimates: All Latent Variable Models



45. Having derived the posterior distributions (comprising 1,000 simulated values) for each model parameter, the next stage is to apply them to ***individuals*** and to then aggregate and summarize the resulting responses at MSOA-level. This is in contrast to the approach used with respect to generating local area estimates from the *2003 Skills for Life Survey*[27] where model parameter posterior distributions were applied to ***area-level data*** describing the aggregate characteristics of those areas. The method ('microsimulation') by which we obtain a population of individual adults in each English MSOA is described on page 42 *et seq.*, but the key here is that by applying the appropriate model parameter posterior distributions to all individuals in all MSOAs (each of whom has their particular characteristics – whether male or female, level of 'highest qualification', tenure, benefit status and so on, including in which MSOA they reside) we generate a corresponding set of 1,000 responses for each MSOA (i.e. of the number of people with each skill level). It is with respect to these sets of 1,000 independent estimates that we then derive summary point estimates and 95% CIs of the number of people with each skill level in each MSOA. Whilst straightforward, this is, as one might imagine, computationally hugely demanding. The purpose, however, is to retain in the final local area estimates all uncertainty that existed in the originating model.

46. It is important here to re-emphasise the essential nature of small area estimation and how this must affect any interpretation of the estimates produced by applying the model parameters to local covariate data. If a particular group of people, nationally, is found to have particularly high skill levels, then it is assumed, unless there is evidence to the contrary, that this will apply to all local areas. The multilevel nature of the model will reveal whether this relationship is mediated by any MSOA-level effects, but the same principle applies – local area estimates are produced on the basis of modelled relationships derived from an analysis of the dataset as a whole.

47. It is thus possible that unknown (and perhaps unknowable) characteristics of one or more particular places have resulted in an anomalous pattern of skills or proportion of ESOL adults – anomalous in the sense that it is quite out-of-line with what might be expected given what is known more widely about the relationship between, on the one hand, individual- and MSOA-level characteristics and, on the other, adult skill levels or ESOL status. Such cannot be captured by small area estimation unless a reasonable sample of individuals is surveyed from all MSOAs in the country, which is clearly not the case with respect to the *2011 Skills for Life Survey* (which comprises a total of 7,230 respondents drawn from 1,516 of the 6,781 MSOAs in England).

48. In fact, a good illustration of this limitation is provided by the fact that, as shown on Figure 16, a small part of rural East Anglia appears to contain improbably high numbers of ESOL adults. The model results in a prediction that over 30% of adults in the Forest Heath 002 MSOA (E02006239) do not speak English as a first language, and that four of the other Forest Heath MSOAs have more than 10% ESOL adults. In fact, Forest Heath is distinctive because of the very high proportion of non-UK born household residents recorded in the *2001 Census* (50.3%). Elsewhere in the country such high percentages are invariably associated with immigrant populations, but in

[27] Gibson, A., Bailey, T, and Fraser, D. (2004) *Demographic mapping of the 2003 Skills for Life Survey to local areas*. Technical Report for the Department for Education and Skills, December 2004.

Forest Heath it is due to the presence of two large American airbases at Lakenheath and Mildenhall. In the *2011 Skills for Life Survey*, as in the country as a whole, being non-UK born is strongly associated with not speaking English as a first language. In Forest Heath this relationship will almost certainly not apply, but without sampling a number of individuals from that particular MSOA (and, in fact, nobody from that MSOA was interviewed as part of the survey) it is impossible for the model to recognise and reflect its uniqueness.[28] The point is that **the results from small area estimation cannot and should not be used as the basis for some sort of 'performance league table' – ranking individual MSOAs, LAs or other geographic or policy units on the basis of, say, predicted adult literacy rates**. The approach is better suited as a mechanism for identifying and highlighting those areas where adult literacy, numeracy and ICT skill levels are *likely* to be relatively poor and where, therefore, it is *likely* that policy initiatives would be best directed. Rare exceptions such as around Lakenheath and Mildenhall notwithstanding, such insights will be generally reliable.

49. In this respect, however, close attention should be paid to estimate precision, which is expressed in terms of Standard Errors or, more commonly, in terms of the range within which we are, given the data, 95% certain that the true value lies (a range obtained directly from the set of prediction estimates). Truly anomalous situations, such as the presence of two large foreign airbases in an otherwise rural part of East Anglia, cannot be accounted for but, by and large, we are able to quantify the level of uncertainty around our estimates. The less information we have that is relevant to a particular place, or the less consistent the relationship between predictor and outcome variables, the less precise will be our estimates – but with sufficient data and good model fit our estimates become more precise. This is why, as illustrated in Figure 24, estimate precision varies between MSOAs; it reflects the evidence that is relevant to each individual MSOA. To take an entirely hypothetical example, if an MSOA comprised an unusually large proportion of Chinese men it is inevitable that the local estimate of skill levels will be very imprecise. There were, after all, only 9 Chinese men interviewed as part of the *2011 Skills for Life Survey* – a poor basis for estimation. But most local areas will not comprise such skewed populations, and the 95% credible intervals around of our estimates shrink accordingly, reflecting the quality of the evidence upon which those estimates are based.

50. That said, there is one important proviso concerning estimate precision, namely that it describes model uncertainty *given the data*. It cannot capture any uncertainty that stems from potential (and unfortunately unquantifiable) inadequacies in the data available to us. The next section thus turns to consider the data that have been used

[28] One potential approach to dealing with this specific problem would have been to use Census data on 'country of birth' in order to distinguish between those born in non-English speaking countries and those born in English speaking countries. Unfortunately, the appropriate table, UV08, has never been produced for MSOAs and thus any use of this data would have required OA to MSOA attribution. More to the point, however, is the fact that this issue only emerged when we examined the mapped output looking for possible anomalies. It was by then simply not possible to repeat the entire analysis, although this is a matter that would need to be addressed in any future small area estimation exercise. The wider point, though, is that it is likely that such anomalies will always emerge and the perspective afforded by small area estimation must always be interpreted in the context of local knowledge that cannot be incorporated within a national-level analysis such as this.

to produce the small area estimates and attempts to evaluate their impact on those estimates.

Skills for Life Small Area Estimation: Data Considerations

51. In essence, our approach to small area estimation rests on the use of two distinct sets of data. On the one hand are the survey data employed, as detailed above, to derive multilevel models which describe how a dependent variable (e.g. literacy skills) responds to a series of individual- and area-level predictor variables. As these data have been collected about adults aged 16-65[29] living in households this defines the population to which we can apply our models. We cannot, in other words, say anything about children or older adults or, importantly, about people living in communal establishments. Any limitations or other issues concerning *SfL Survey* data will have been considered elsewhere,[30] though it is worth noting that Small Area Estimation is not as sensitive to sampling issues as traditional survey-based estimation methodologies. Obtaining as diverse a sample population as possible – given sample size – is the key criterion and, in this respect, the *SfL Survey* provides an entirely adequate basis for modelling.

52. On the other hand are those data used to define the socio-demographic composition of local areas, along with area-level variables that equate to those used in the upper-level of the model. The survey and local covariate data must correspond as the goal is to apply posterior distributions for the various factors used in the model to *individuals* in local areas. Thus, to take a hypothetical (and unrealistically simple) example, if a model included individual-level terms for three age bands and one sex, along with a single upper-level variable, say the *2010 Index of Deprivation*, then it is necessary to establish how many people in each area are in each of the 8 age-sex categories (which allows for reference groups), as well as each area's *2010 IMD* score. It is to individuals (with their age and sex characteristics) in areas (with their *2010 IMD* scores) that the modelled posterior distributions are applied in order to establish the likelihood that those individuals will have a particular skill level. By summing those likelihoods across each MSOA as a whole we derive estimates of the number of people in each skill level in each MSOA.[31]

53. This may appear relatively straightforward, but great complexity arises once 'real-world' models are constructed. Take, for example, the literacy model described in Table 3. This comprises seven individual-level variables: benefit status (2 categories); economic activity status (4 categories); occupation (10); tenure (3); limiting long-term illness (2); age band (4) and sex (2 categories). It defines, in other words, some 2 × 4 × 10 × 3 × 2 × 4 × 2 = 3,840 different 'person types'. The other

---

[29] Note that, although the *SfL Survey* samples 16-65 year olds, we can only make predictions for 16-64 year olds as this is the age-banding used by the 2001 Census and DWP returns.

[30] Harding, C., et al (forthcoming) 2011 Skills for Life survey: a survey of literacy, numeracy and ICT levels in England. Department for Business, Innovation and Skills Research Paper. Annex 1.

[31] It is worth noting that a different approach was followed in the *Demographic mapping of the 2003 Skills for Life Survey to local areas* project. In that project, rather than applying model factor posterior distributions to individuals according to their individual-level characteristics, the posterior distributions were applied to local areas (2003 statistical wards, which are precisely equal to the Standard Table (ST) wards used in this report), weighted according to the socio-demographic composition of those local areas. Whilst far less technically demanding (not least because it does not require each and every local area to be microsimulated – see below) this is a far less theoretically-satisfying methodology than that now being used because information is lost when using aggregated local area data.

skills models, described in Table 4 to Table 8, incorporate one or more additional individual-level variables and define 34,560 person types (in the case of the numeracy model), 69,120 person types (the email and word processing models), and 138,240 person types (for the spreadsheet and ICT multiple choice models). The ESOL model is, by comparison, relatively undemanding, though even here a total of 8,640 person types are defined. The challenge, then, is that it is necessary to determine how many of each person type there are in each of the 6,781 MSOAs for which predictions are required.

Microsimulation

54. Unfortunately, information concerning the detailed composition of MSOA populations simply does not exist. It is, however, possible to use what is known about the aggregate characteristics of any given population (i.e. how many males and females, how many people with or without a limiting long-term illness, how many people in each age band, etc.) in order to deduce – or *microsimulate* – the likely number of people with each unique combination of characteristics (for instance the number of 16-34 year-old males with and without a limiting long-term illnesses; the number of 16-34 year-old females with and without limiting long-term illnesses; and so on).[32] The defining characteristic of a successfully microsimulated population is that, when aggregated, it will match in all respects what is known about the overall characteristics of that population.

55. The 2001 Census provides, as listed in Table 12 below, a number of univariate and multivariate tables which can be used to derive information on the aggregate characteristics of MSOA populations (the so-called 'marginal distributions' for what is an unknown 'full joint distribution'), as well as on a number of 'partial' joint distributions; such as CAS017 which cross-tabulates long-term limiting illness with tenure, or CAS021, which describes a three-way joint distribution comprising sex, long-term limiting illness and economic activity. Additional information on the number of people in MSOAs on Incapacity Benefit and Severe Disablement Allowance (used in all models) is derived from LSOA-level 4th Quarter 2010 data in the *Work and Pensions Longitudinal Study* dataset.[33]

[32] Ballas, D., Dorling, D., Thomas, B., & Rossiter, D. (2005) *Geography matters: simulating the local impacts of national social policies*. Joseph Roundtree Foundation. doi:10.2307/3650139 (Available online at http://www.jrf.org.uk/publications/geography-matters-simulating-local-impacts-national-social-policies.) [Accessed 16/1/2012.]

[33] The WPLS dataset, from which individual quarterly data can be extracted, is available via NOMIS (http://www.nomis.co.uk). [Accessed 20/4/2012].

Table 12 2001 Census Tables used in the microsimulation of MSOA Populations

Census Area Statistics (CAS) Tables	
CAS001	Age by sex and whether living in a household or communal establishment
CAS016	Sex and age by general health and limiting long-term illness
CAS017	Tenure and age by general health and limiting long-term illness
CAS021	Economic activity by sex and limiting long-term illness
CAS026	Sex and economic activity by general health and provision of unpaid care
CAS032	Sex and age and level of qualifications by economic activity
CAS033	Sex and occupation by age
CAS034	Former occupation by age
CAS061	Tenure and car or van availability by economic activity
CAS105	Age by highest level of qualification
CAS113	Occupation by highest level of qualification
Census Area Statistics Theme (CAST) Tables	
CAST03	Theme table: ethnic group cross-tabulated by (a) sex; (b) ageband; (c) birthplace; (d) economic activity; (e) limiting long-term illness; and (f) resident type.
Key Statistics (KS) Tables	
KS05	Country of birth
Univariate (UV) Tables	
UV03	Sex
UV04	Age
UV09	Ethnic group (England and Wales)
UV22	Limiting long-term illness
UV24	Qualifications (England and Wales)
UV28	Economic activity
UV30	Occupation
UV43	Tenure (people) (England, Wales and Northern Ireland)
UV71	Communal establishment residents

56. A technique known as 'Iterative Proportional Fitting' (IPF) has long been used as a method of combining marginal distributions (and two- and three-way joint distributions) to derive a full joint distribution – i.e. one which describes the number of individuals in a population with each unique combination of characteristics.[34] We

[34] Deming, W.E. and Stephan, F.F., 1940, "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known", *Annals of Mathematical Statistics*, Vol. 11, pp427-444; Fienberg, S.E., 1970, "An iterative procedure for estimation in contingency tables", *Annals of Mathematical Statistics*, Vol. 41, pp907-917; Clarke, M. and Holm, E., 1987, "Microsimulation methods in human geography and planning: a review and further extensions", *Geografiska Annaler*, Vol. 69B, pp145-164; Birikin, M. and Clarke, M., 1988, "SYNTHESIS – a synthetic spatial information system for urban and regional analysis: methods and examples", *Environment and planning A*, Vol. 20, pp1645-1671.

have built upon this approach to develop a method which can cope with the fact that individual cells in census tables have been perturbed by a variety of disclosure control methods. What this means is that cells will not necessarily aggregate to the same partial or marginal distributions. Consider, for instance, aspects of MSOA E02000001 (City of London) as revealed by a variety of different tables. As shown in Table 13 below, even the total population recorded by the census varies from 7,154 to 7,234 depending on which sets of table cells are aggregated. Sub-aggregations similarly vary with, for instance, the number of people aged 16-74 in employment varying between 4,242 and 4,290 depending on the table used. The point is that, because of disclosure control, the census tables do not provide a straightforward set of marginal and 2- and 3-way tables describing a fixed, if unknown, full joint distribution.[35]

Table 13 Consequences of Disclosure Control: Summations of Census data for the City of London (MSOA E02000001)

All people, all ages	
CAS001	7,199
CAST03a	7,234
CAST03b	7,154
CAST03c	7,190
CAST03d	7,165
CAST03e	7,198
CAST03f	7,164
KS05	7,185
UV03	7,185
UV04	7,187
UV09	7,185
UV22	7,185
UV71	7,187

People in Households, all ages	
CAS016	6,819
CAS017	6,828
UV43	6,861

All people, aged 16-74	
CAS021	6,061
CAS032	6,064
CAS105	6,085
CAS113	6,035
UV24	6,067
UV28	6,067

All people, 16-74, in employment	
UV30	4,290
CAS033	4,242

57. We have therefore developed an approach which seeks to iteratively assign people to 'cells' in a simulated full joint distribution in such a way as to minimise a test statistic which sums the aggregate difference between the resulting marginal estimates and the 'known', but conflicting, marginal totals taken from the census tables listed in Table 12 above and the DWP *Work and Pensions Longitudinal Study* dataset.

[35] Part of the problem lies with the fact that CAS and CAST tables are only available at Output Area (OA) level and have had to be aggregated up to MSOA level. Disclosure control affects the individual OA tables and, through aggregation, one can compound the uncertainty inherent in the census data. This cannot be avoided if one wishes to take advantage of the 2- and 3-way partial joint distributions offered by these tables.

58. At the heart of this approach is a (0,1) relational matrix linking each person type with each of 584 'known' marginal totals. The simplified exemplification given in Table 14 below illustrates how this matrix allows us to relate the number of each person type to the total number who would thus appear in each *estimated* marginal total.[36] The goal is to determine what composition of person types would result in a set of *estimated* marginal totals that equals, or very closely approximates, the set of *known* marginal totals.

Table 14 Calculating marginal total estimates using a link matrix

	Link Matrix						Estimated count of Person Type
Person Type	**M**	**F**	**16-54**	**55-74**	**LLTI**	**No LLTI**	
M, 16-54, LLTI	1	0	1	0	1	0	← 27
M, 16-54, No LLTI	1	0	1	0	0	1	← 14
M, 55-74, LLTI	1	0	0	1	1	0	← 7
M, 55-74, No LLTI	1	0	0	1	0	1	← 5
F, 16-54, LLTI	0	1	1	0	1	0	← 14
F, 16-54, No LLTI	0	1	1	0	0	1	← 16
F, 55-74, LLTI	0	1	0	1	1	0	← 8
F, 55-74, No LLTI	0	1	0	1	0	1	← 9
	↓	↓	↓	↓	↓	↓	
Estimated Marginal	53	47	71	29	56	44	Sum of absolute differences
Known Marginal	51	49	75	25	53	47	
Absolute Difference	2	2	4	4	3	3	→ 18

59. In this case, which has not been subjected to any disclosure control perturbation, a unique solution is possible (shown in Table 15 below). In reality, not only is the size of the matrix very much greater (there are, depending on the model, between 3,840 and 138,240 person types, rows, and 584 marginal totals, columns), but there is no actual solution because, as described above, the *known* marginal totals conflict. Our method was thus, for each MSOA in turn, to;

    a) Set the initial estimate of the count of each person type to equate to the composition of the 2001 Census Individual Sample of Anonymised Records[37] of the region in which the MSOA lies, weighted so that the total number of people equals the known MSOA population (which is taken from Table UV03).

    b) Calculate the test statistic as $T = \sum_{i=1}^{584} \left| \text{KnownMarginal}_i - \text{EstimatedMarginal}_i \right|$

[36] This hypothetical table is entirely invented, although limiting long-term illness (LLTI) is described in the Glossary.

[37] Office for National Statistics, *2001 Census: Sample of Anonymised Records (SARs)* (Licensed) (England, Wales, Scotland and Northern Ireland) [computer file]. ESRC/JISC Census Programme, Cathie Marsh Centre for Census and Survey Research (University of Manchester). See http://www.census.ac.uk/guides/Microdata.aspx [Accessed 18/1/2012.]

c) Move individuals between person types to reduce the test statistic T (using an algorithm to determine the most effective cell-to-cell swaps)

d) Continue iterating until the test statistic converges to a steady state.

60. For convenience (rather than efficiency) this was implemented in R. The resulting 'microsimulated' MSOA populations do not sum to precisely match any individual census table's marginal totals, but they do offer a 'best-fit' which averages out the effects of disclosure control on individual tables. Convergence in all cases resulted in a T-statistic of less than 5,000[38]. Analysis of the microsimulated populations shows that the method discriminates very well between MSOAs, with the estimated marginal totals for all MSOAs always being far closer to their own 'known' marginal totals than to the known totals for any other MSOA.

Table 15 A 'solved' illustrative microsimulated population

	Link Matrix						Estimated count of Person Type
Person Type	M	F	16-54	55-74	LLTI	No LLTI	
M, 16-54, LLTI	1	0	1	0	1	0	← 26
M, 16-54, No LLTI	1	0	1	0	0	1	← 15
M, 55-74, LLTI	1	0	0	1	1	0	← 5
M, 55-74, No LLTI	1	0	0	1	0	1	← 5
F, 16-54, LLTI	0	1	1	0	1	0	← 16
F, 16-54, No LLTI	0	1	1	0	0	1	← 18
F, 55-74, LLTI	0	1	0	1	1	0	← 6
F, 55-74, No LLTI	0	1	0	1	0	1	← 9
	↓	↓	↓	↓	↓	↓	
Estimated Marginal	51	49	75	25	53	47	Sum of absolute differences
Known Marginal	51	49	75	25	53	47	
Absolute Difference	9	0	0	0	0	0	→ 0

61. Due to the disclosure control measures adopted in the *2001 Census*, microsimulation cannot precisely reconstitute the socio-demographic composition of individual MSOA populations, but there is no question that it provides a close approximation. More problematic is the simple and unavoidable fact that the *2001 Census* was undertaken the best part of a decade before the *2011 Skills for Life Survey*. This may not matter everywhere, but in some areas there will undoubtedly have been significant changes to the socio-economic composition of populations. These cannot be captured as, until the detailed results of the *2011 Census* are published in 2012/13, there is no other source of comparable small area socio-economic data available for England.

62. Demographic changes since the *2001 Census* can, however, be reflected in local area estimates by adjusting the estimates based on the *2001 Census* so that they

[38] This means that across all MSOAs known and estimated marginal totals differed, on average, by less than 5 people.

align with the most recent (2009) ONS mid-year age-sex population estimates available at the time of analysis.[39] Whilst these are deemed by the ONS to be 'experimental' statistics, they provide the only official source of data on the demographic composition of contemporary populations. As these data refer to all people (i.e. those in households and in communal establishments) we have, for each age-sex category in each MSOA, derived a factor which describes how that specific population has changed since 2001. We essentially perform this by simple raking[40], keeping proportions within an age-sex group constant (e.g. the proportion of an age sex band with a given social status, qualification, economic activity etc.) but matching the eight age-sex totals for each MSOA to the 2009 estimates rather than the 2001 census. The resulting estimates are given in the *middle-super-output-areas-2009-all.xlsx* and *middle-super-output-areas-2009-el-l1.xlsx* local area predictions files. We must presume, in passing, that factors derived from an analysis of *overall* population changes for males and females in the 16-34, 35-49, 50-59 and 60-64 agebands can be legitimately applied to adults in households; which is the target population for whom estimates are being modelled.

63. This is a straightforward process. Thus if we had predicted that 100 males aged 35-49 in a given MSOA would have, say Entry Level 1 literacy, then, if that age-sex group for that MSOA had increased in size by 10% between 2001 and 2009, we would adjust our estimate to 110 people. All other estimates would be similarly adjusted, so that the sum of the estimates matches the total number of 35-49 males in that MSOA. The problem, of course, is that this presumes that there has been no change in the socio-economic characteristics of the population cohort.

64. By and large, with relatively static housing stock, this is likely to be a reasonable assumption, but in some cases the scale of change must raise serious concerns about the reliability of local area predictions. The number of 60-64 females in the Swindon 002 MSOA (E02003213) increased, for instance, by over 540% between 2001 and 2009. This must reflect new housing which may well have resulted not just in a different demographic structure but in a quite different socio-economic profile. In fact, the Swindon 002 MSOA saw its overall population nearly triple in size between 2001 and 2009, and it seems inconceivable that this will not have affected its socio-economic composition in ways that would also have affected the distribution of skills (and proportion of ESOL adults) in the area.

65. The fact of the matter is that the principal weakness in our small area estimates lies with the fact that ***we must make the crucial assumption that all areas retain the same socio-economic composition as they had in 2001, and have the same demographic composition as estimated for 2009*** (an ONS estimate which may, or may not, be well-founded). To a degree this shortcoming can be mitigated by applying local knowledge when interpreting the small area estimates, but it is nevertheless unfortunate that the results of the *2011 Census* were not available at the time of analysis. These are due to be published over the course of 2012/2013

---

[39] Office for National Statistics, *Super Output Area mid-year population estimates for England and Wales, Mid-2009*. Available at http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-213438 [Accessed 20/4/2012.].

[40] For a description of raking, see http://www.quantitativeskills.com/sisa/calculations/rakehlp.htm. [Accessed 20/4/2012.]

and they will provide an up-to-date, detailed and largely reliable picture of the socio-demographic structure of contemporary England. If Small Area Estimates are to inform policy we must recommend in the strongest possible terms that the exercise is repeated using data drawn from the *2011 Census*. Given our understanding that this will be very similar in structure and content to the *2001* Census, this would not entail any additional modelling, but rather just the re-microsimulation of MSOA populations (using 2011 census data) and the application of existing model posterior distributions to the new covariate data. It would also require the attribution of modelled MSOA-level estimates to the other geographies of interest, as described below.

Attributing to other Geographies

66. Middle Layer Super Output Areas (MOSAs) have become the *de facto* standard geography for which administrative data is published. As a result, it is for MSOAs that we can obtain the most suitable and up-to-date covariate data and for which, as a consequence, we have produced modelled estimates – both on the basis of 2001 populations and, as described above, for 2001 populations weighted to align with the ONS's 2009 age-sex population estimates. Both sets of estimates have been made available, and maps illustrating the '2009 MSOA' estimates are presented in Section 0. Yet estimates are also required for a number of other geographies for which little or no relevant covariate data is available. For these geographies it has been necessary to attribute MSOA-level estimates on the basis of the February 2011 *Open National Statistics Postcode Directory* (*ONSPD*).[41]

67. The ONSPD provides a count of residential addresses in each postcode, and lists within which higher geographies – MSOAs, Local Authorities, Parliamentary Constituencies, etc. – these postcodes lie. On this basis it possible to derive address-weighted lookup tables relating MSOAs (n=6,781) to each of the other geographies of interest: namely Standard Table (ST) wards (n=7,932); 2005 Statistical wards (n=7,972); 2011 Council wards (n=7,618); 2011 Parliamentary Constituencies (n=533); Local Authorities (n=326); and Local Enterprise Partnership areas (n=37).[42] These lookup tables are used to allocate the 2009 MSOA-level posterior estimates as appropriate, and mean estimates and 95% CIs for the new geographies are derived directly from the re-distributed and re-aggregated sets of posterior estimates. Thus, if a 2011 Council ward comprised half the addresses of three different MSOAs, then 50% of each posterior estimate of the number of people with, say Level 1 literacy, would allocated to that ward. The result would be, for that ward, a corresponding set of posterior estimates upon which to calculate the mean and 95% CIs. In this way we once again seek to retain full information concerning model uncertainty as it affects the estimates for the new geographies.

68. The particular issue here, however, is that, because of variations in household size and composition, this use of addresses will provide only an imperfect guide to the distribution of adult residential populations between different units. Because of the

---

[41] ONS, *Office for National Statistics Postcode Directory (ONSPD)* Open February 2011 edition, 2011. Available at http://www.sharegeo.ac.uk/handle/10672/203. [Accessed 12/1/2012.]

[42] The recent addition of (a) the Northamptonshire and (b) Buckinghamshire Thames Valley Local Enterprise Partnerships has increased the total to 38 (as of April 2012), but these were announced too late to be included in the present analysis. See footnote 3 above.

degree of spatial overlap vis-à-vis MSOAs, this is of particular concern with respect to the composition of ward populations. As a guide to local variations in adult skills and ESOL status, ward estimates must therefore be considered secondary to those produced for MSOAs. They may be useful if they are to be set against comparative data only available at ward level, but we would otherwise recommend against their use. Indeed, not only are these ward-level estimates derived from estimates generated using MSOA-level data, but the wards themselves, unlike MSOAs, were not designed for statistical analysis. For instance, all three ward geographies include a number of City of London wards with very small residential populations and there is one 2011 Council ward (30UHHH) with no residential addresses listed in the ONSPD. This, in fact, refers to the campus of Lancaster University which contains almost no household population and, as a result, fell below the threshold for 2001 census ward output (being amalgamated with 30UHGN, Ellel ward). At the other end of the scale, there are some wards containing over 30,000 people.

69. MSOAs, by contrast, were developed by Neighbourhood Statistics as part of a hierarchy of units specifically designed for the collection and publication of small area statistics.[43] They are of much more consistent size and are not subject to the frequent boundary changes that affect all types of ward (indeed, statistical wards are no longer maintained, the last 'edition' being the 2005 Statistical wards for which we have generated small area estimates). They were, moreover, generated by zone-design software which automatically grouped together 2001 census output areas (OAs) into Lower Layer Super Output Areas (LSOAs), and LSOAs into MSOAs, according to a range of designated size, boundary and 'homogeneity' criteria. They have since become the standard unit for geographic analysis and most data are now made available at MSOA level. For these reasons, as well as the fact that it is for MSOAs that we provide 'direct' estimates, ***we strongly recommend that MSOAs, rather than any of the ward geographies, are used for any subsequent policy or analytical purposes***.

70. Local Authorities, Parliamentary Constituencies, and Local Enterprise Partnership areas are, some boundary discontinuities notwithstanding, largely aggregations of MSOAs and are not subject to the same limitations as the smaller ward geographies.

[43] See the discussion regarding the design of lower and middle layer super output areas (MSOAs) on the ONS's Neighbourhood Statistics website: http://www.neighbourhood.statistics.gov.uk/dissemination/Info.do?page=aboutneighbourhood/geography/superoutputareas/soafaq/soa-faq.htm. [Accessed 12/1/2012.]

Summary of Results

71. Whilst this study is not concerned with an analysis of the small area estimates themselves – or what they may tell us about the distribution of adult skills deficits (and of ESOL adults) – it is important to recognise the extent to which the results exhibit 'face validity'. The degree to which, in other words, the results are both internally consistent and align with what might be expected.

72. In large measure, this assessment of face validity will have to lie with those expert in the field, but we can here attempt to summarise the findings and draw attention to a number of key features. To that end this final section includes a series of maps plotting the estimates at the smallest scale – MSOAs (Figure 10 to Figure 16) – as well as at the local authority scale (Figure 17 to Figure 23). These maps aim to divide the data into broadly similar groups for display purposes, as detailed below. The maps have been designed to emphasise areas of skill deficits *relative* to the underlying norm for that particular skill domain; i.e. areas are most strongly shaded where there is a much higher than average percentage of 'adults in households' with Entry Level or below skills.

Table 16 MSOA Maps: Choropleth Classification of Entry Level and Below Skills

Literacy	Lowest Group	Low Mid Group	Middle Group	High Mid Group	Highest Group
Group Range	<14%	14%-16%	16%-18%	18%-20%	>20%
(% of MSOAs)	(38.8%)	(22.9%)	(14.3%)	(9.6%)	(14.3%)
Numeracy					
Group Range	<44%	44%-49%	49%-54%	54%-59%	>59%
(% of MSOAs)	(32.2%)	(23.8%)	(19.0%)	(12.4%)	(12.6%)
ICT - Email					
Group Range	<44%	44%-48%	48%-52%	52%-56%	>56%
(% of MSOAs)	(39.2%)	(18.1%)	(16.5%)	(11.9%)	(14.3%)
ICT – WordPro					
Group Range	<65%	65%-69%	69%-72%	72%-75%	>75%
(% of MSOAs)	(38.3%)	(20.0%)	(14.0%)	(11.6%)	(16.1%)
ICT - Spreadsheets					
Group Range	<69%	69%-72%	72%-75%	75%-78%	>78%
(% of MSOAs)	(35.8%)	(17.2%)	(17.0%)	(13.3%)	(16.6%)
ICT - Multiple Choice					
Group Range	<22%	22%-26%	26%-30%	30%-34%	>34%
(% of MSOAs)	(32.4%)	(21.7%)	(17.3%)	(12.1%)	(16.5%)
ESOL					
Group Range	<2%	2%-3%	3%-5%	5%-9%	>9%
(% of MSOAs)	(32.3%)	(19.9%)	(18.0%)	(11.8%)	(18.0%)

Table 17 LA Maps: Choropleth Classification of Entry Level and Below Skills

Literacy	Lowest Group	Low Mid Group	Middle Group	High Mid Group	Highest Group
Quintile Range	<15%	15%-16%	16%-17%	17%-19%	>19%
% of LAs	42.5%	17.1%	16.5%	16.2%	7.6%
Numeracy					
Quintile Range	<46%	46%-49%	49%-52%	52%-55%	>55%
% of LAs	33.0%	22.9%	20.2%	15.3%	8.6%
ICT - Email					
Quintile Range	<46%	46%-48%	48%-50.5%	50.5%-54%	>54%
% of LAs	42.2%	16.2%	17.7%	15.9%	8.0%
ICT - WordPro					
Quintile Range	<66%	66%-68.5%	68.5%-71%	71%-74%	>74%
% of LAs	37.0%	18.7%	19.0%	17.7%	7.6%
ICT - Spreadsheets					
Quintile Range	<70%	70%-72%	72%-74%	74%-77%	>77%
% of LAs	33.0%	18.0%	18.7%	20.8%	9.5%
ICT - Multiple Choice					
Quintile Range	<23%	23%-26%	26%-29%	29%-32%	>32%
% of LAs	25.4%	28.1%	20.8%	18.0%	7.6%
ESOL					
Quintile Range	<2.5%	2.5%-3.5%	3.5%-5%	5%-15%	>15%
% of LAs	35.8%	19.6%	16.2%	20.5%	8.0%

73. The MSOA maps exhibit impressive granularity, though it worth repeating the observations made above about the nature of these *modelled* estimates. Small Area Estimation 'pools' evidence to 'enhance' local estimates; generating, in other words, local area estimates on the basis of modelled relationships that have been derived from an analysis of the dataset as a whole. It is best suited, therefore, as a mechanism for identifying those areas where adult literacy, numeracy and ICT skill levels are *likely* to be relatively poor and where, therefore, it is *likely* that policy initiatives would be best directed. It cannot, as already illustrated with respect to the impact of the USAF airbases at Mildenhall and Lakenheath in Suffolk, possibly hope to capture genuinely anomalous variations from the norm. After all, we are here seeking to estimate adult skills levels etc. in 6,781 MSOAs (and in significantly more Standard Table, Statistical and Council wards) on the basis of a national survey of just 7,230 adults.

74. Yet, based on a set of multilevel models, the analysis has, as illustrated by Figure 10 to Figure 16, served to highlight some important patterns with respect to adult skills at the local level. First, as might be expected, adult skill levels show huge variations at the very local level. For instance, although the proportion of people with Entry Level or below literacy varies from 11.3% (Wokingham) to 21.6% (Knowsley) at local authority level, at MSOA-level the proportion ranges from 8.8% (Basingstoke & Deane 021) to 33.6% (Liverpool 022). This pattern is repeated across the numeracy and ICT skills domains, and it is clear that areas of adult skill deficits can be very localised. Small Area Estimation provides an invaluable means of highlighting and quantifying this local dimension.

75. Second, and continuing this theme, it is clear that there are some very marked variations in skill levels within individual local authority areas. Unsurprisingly, such variation tends to be most marked in large urban authorities, such as Liverpool which, although it has an overall literacy skills deficit rate of 20.8% (i.e. of people with Entry Level or below literacy skills), includes individual MSOAs with rates varying from 11.6% to 33.6%. Rural areas are not exempt from this, with many rural local authorities showing similar diversity; such as in North East Lincolnshire which, although having an overall rate of 17.3%, contains individual MSOAs with rates which vary from 13.7% to 28.3%. The perspective afforded by Small Area Estimation strongly suggests that any policy response to skills deficits will need to be spatially fine grained.

76. Third, the detailed MSOA maps (along, indeed, with the LA-level maps) draw attention to the fact that adult skills deficits are not a purely urban phenomenon, although the densest concentrations of poor literacy, numeracy and ICT skills do appear to affect England's principal conurbations. This rural dimension appears particularly marked across the ICT skills domains, with relatively high rates of Entry Level and below skill levels emerging in many south western and northern counties, as well as around the Wash. Once again, in other words, the evidence afforded by Small Area Estimation adds significantly to our understanding of the distribution and scale of adult skills deficits in England.

77. Turning to the distribution of ESOL adults, here the pattern is very much as might be expected – excepting, of course, the Mildenhall and Lakenheath effect in East Anglia. Other such anomalous results may be embedded within Figure 16 and be apparent to those with a detailed knowledge of particular areas but, by and large, both the overall structure (which exhibits the classic 'Bristol Channel to the Wash' division of England) and more local patterns appear uncontroversial. The presence of high rates of ESOL adults across all of London except in those traditional 'working-class' communities that border Essex, as well as the more affluent boroughs in the south east of the city, is striking; as is the spatial clustering of ESOL adults in the old cotton towns of Lancashire. More generally, Figure 16 illustrates particularly clearly the concentration of ESOL adults in the centres of many large towns and cities across England.

78. Our final comment should perhaps be directed at the issue of estimate uncertainty. Developing a methodology that could capture estimate uncertainty was a key design consideration for this project, but it has become apparent that what we have termed 'model uncertainty' – i.e. the quantifiable level of confidence we have in our modelled estimates *given the data* – is likely to significantly underestimate the overall level of *model+data* uncertainty. We include in the output files (described in Section 0 below) the 95% CIs for each and every estimate of the number and proportion of people in each area with each literacy, numeracy and ICT skill level, as well as for all estimates of the number and proportion of ESOL adults in each area, but these are estimates of model uncertainty only and must thus be treated with some caution.

79. Figure 24 on page 68 below illustrates graphically the relative size of the 95% CIs around the 6,781 MSOA-level literacy estimates; focussing here on the proportion of people with Level 2 or above literacy skills. An additional, and quite unquantifiable, degree of uncertainty surrounds these relatively narrow CIs simply because we cannot know the extent to which our understanding of the socio-economic

composition of local populations – which rests largely on evidence drawn from the *2001 Census* – reflects current realities on the ground. In many parts of the country there will have been relatively little change since 2001 – and we may have considerable confidence in our estimates – but in some places, and we cannot know where, those changes are likely to have been significant and are likely to have affected both levels of adult skills and the proportion of people who do not speak English as a first language. The additional (and unquantifiable) uncertainty stems from the fact that we simply do not know where, or how, populations have changed over the past decade. Only by repeating this analysis using *2011 Census* data would it be possible to obtain a fully-satisfactory insight into local patterns of adult literacy, numeracy and ICT Skills, as well as the local distribution of ESOL adults.

Figure 10 Entry Level and Below Literacy Skills: English MSOAs (n=6,781)

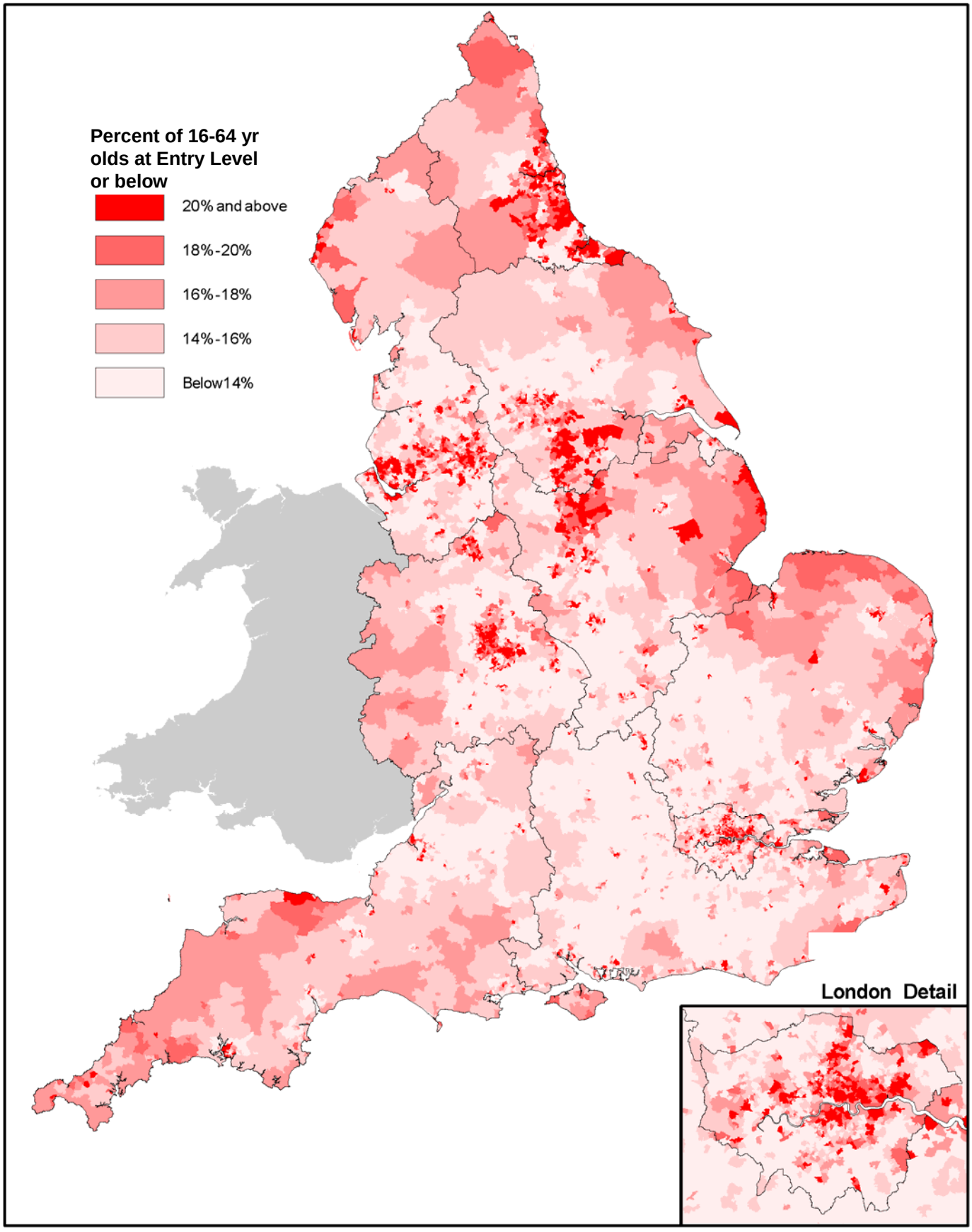


Figure 11 Entry Level and Below Numeracy Skills: English MSOAs (n=6,781)

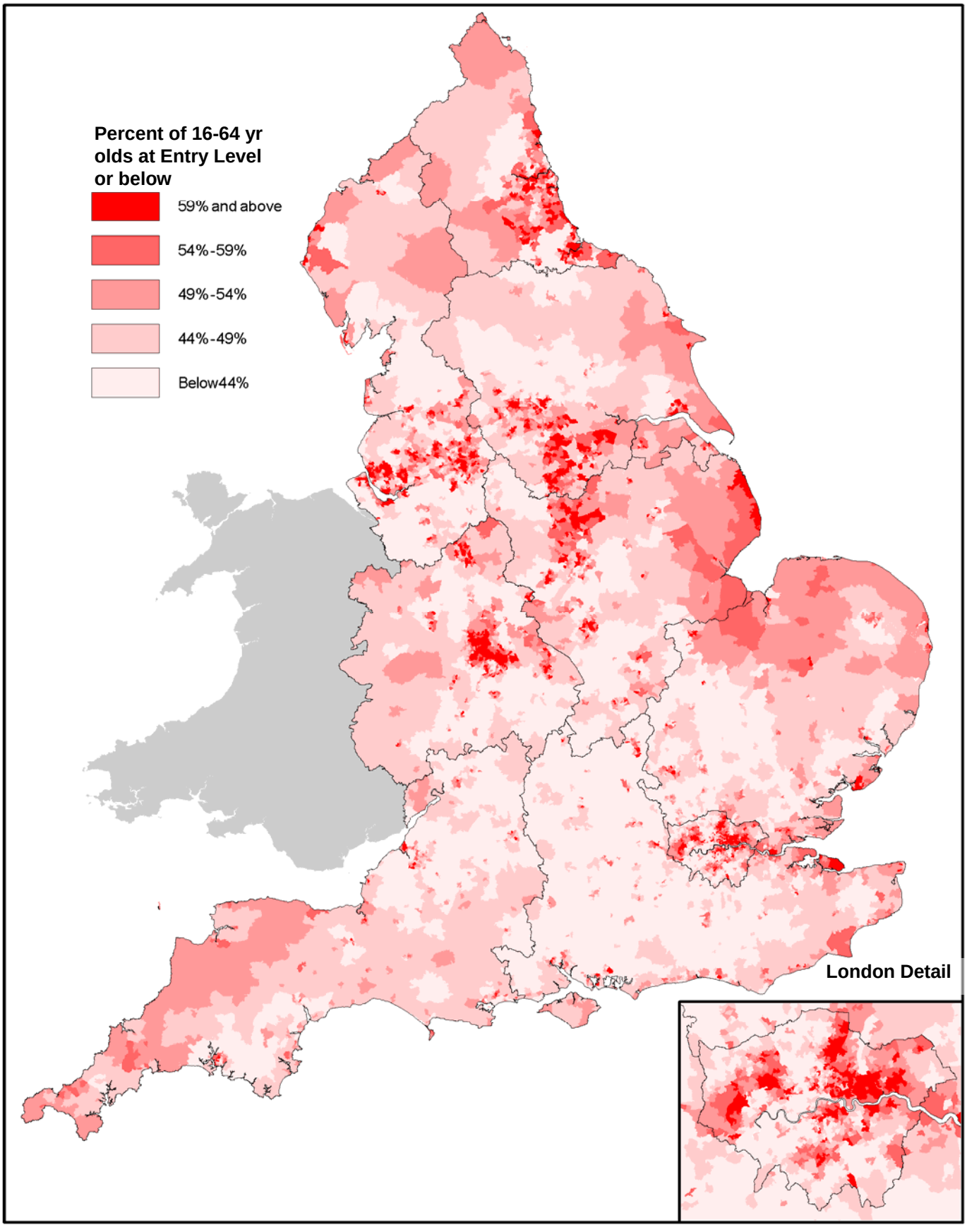


Figure 12 Entry Level and Below Email Skills: English MSOAs (n=6,781)

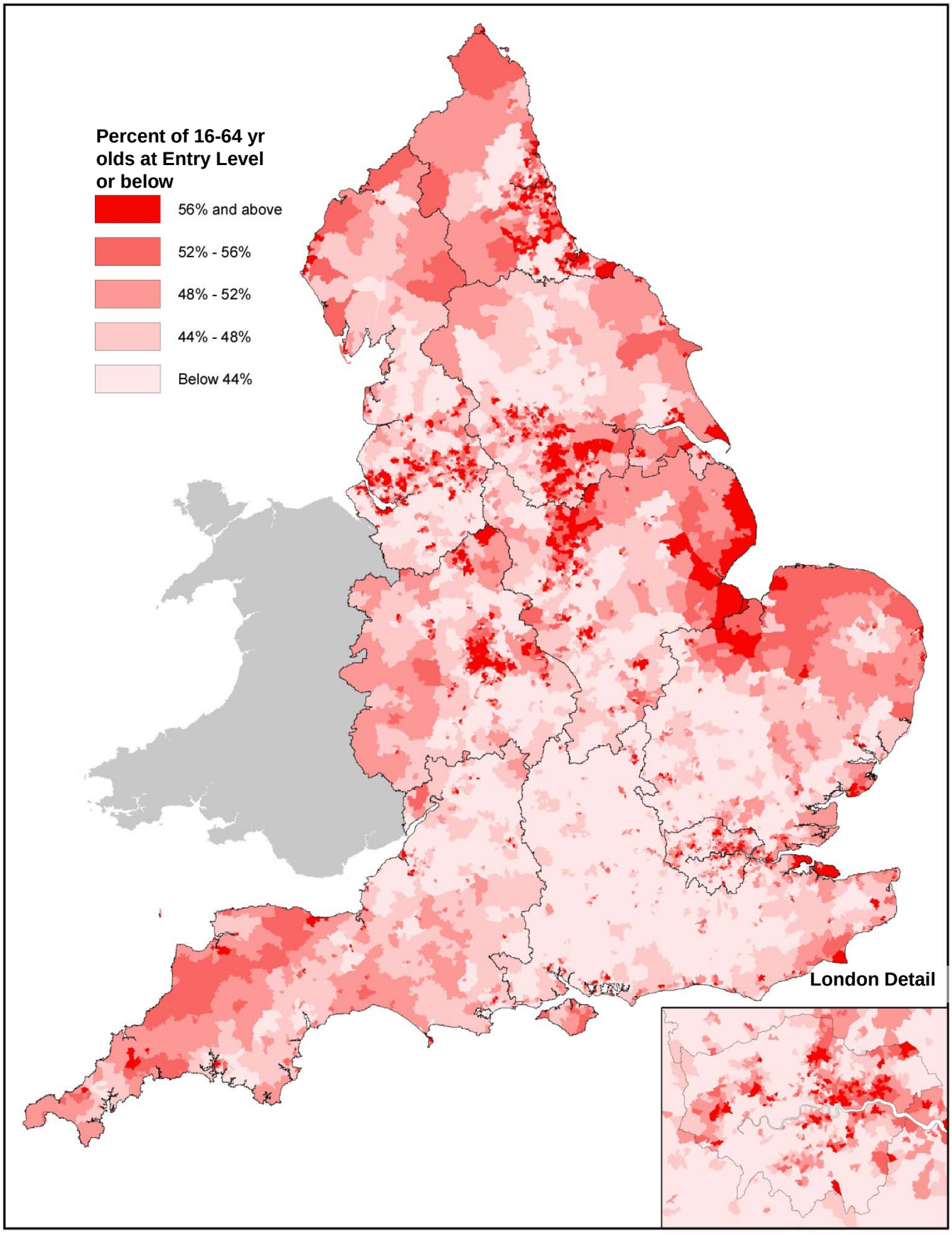


Figure 13 Entry Level and Below Word Processing Skills: English MSOAs (n=6,781)

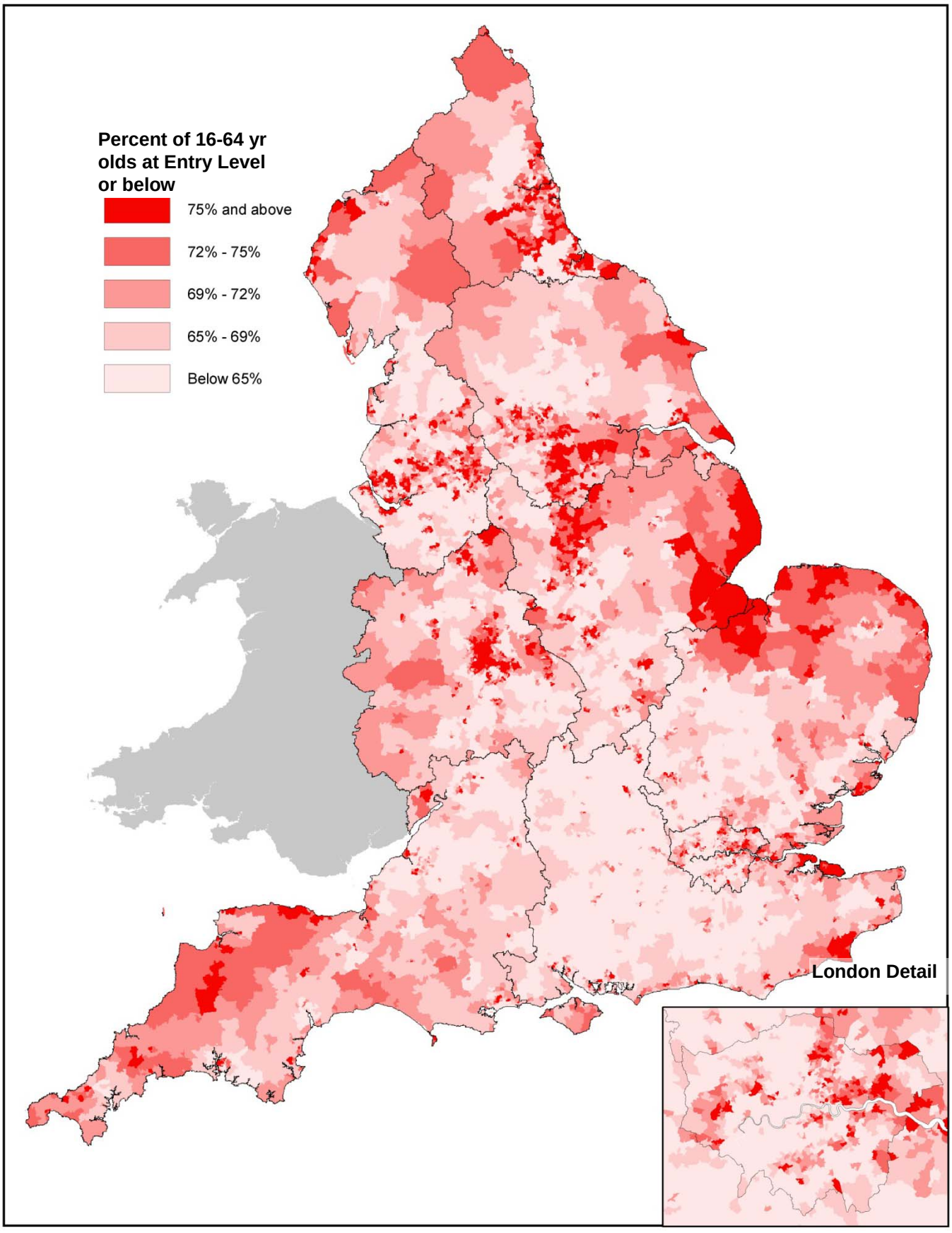


Figure 14 Entry Level and Below Spreadsheet Skills: English MSOAs (n=6,781)

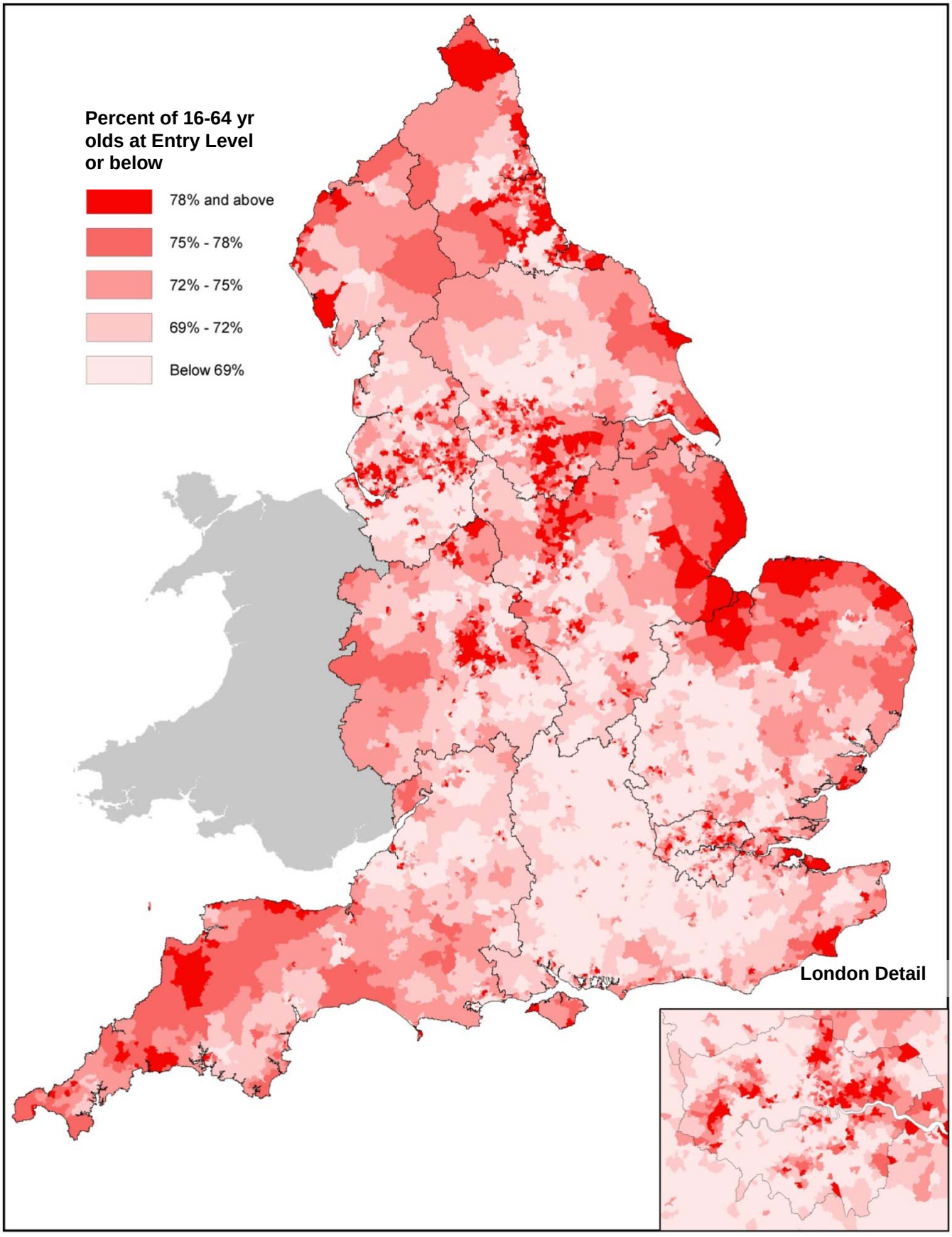


Figure 15 Entry Level and Below ICT (Multiple Choice Test) Skills: English MSOAs (n=6,781)

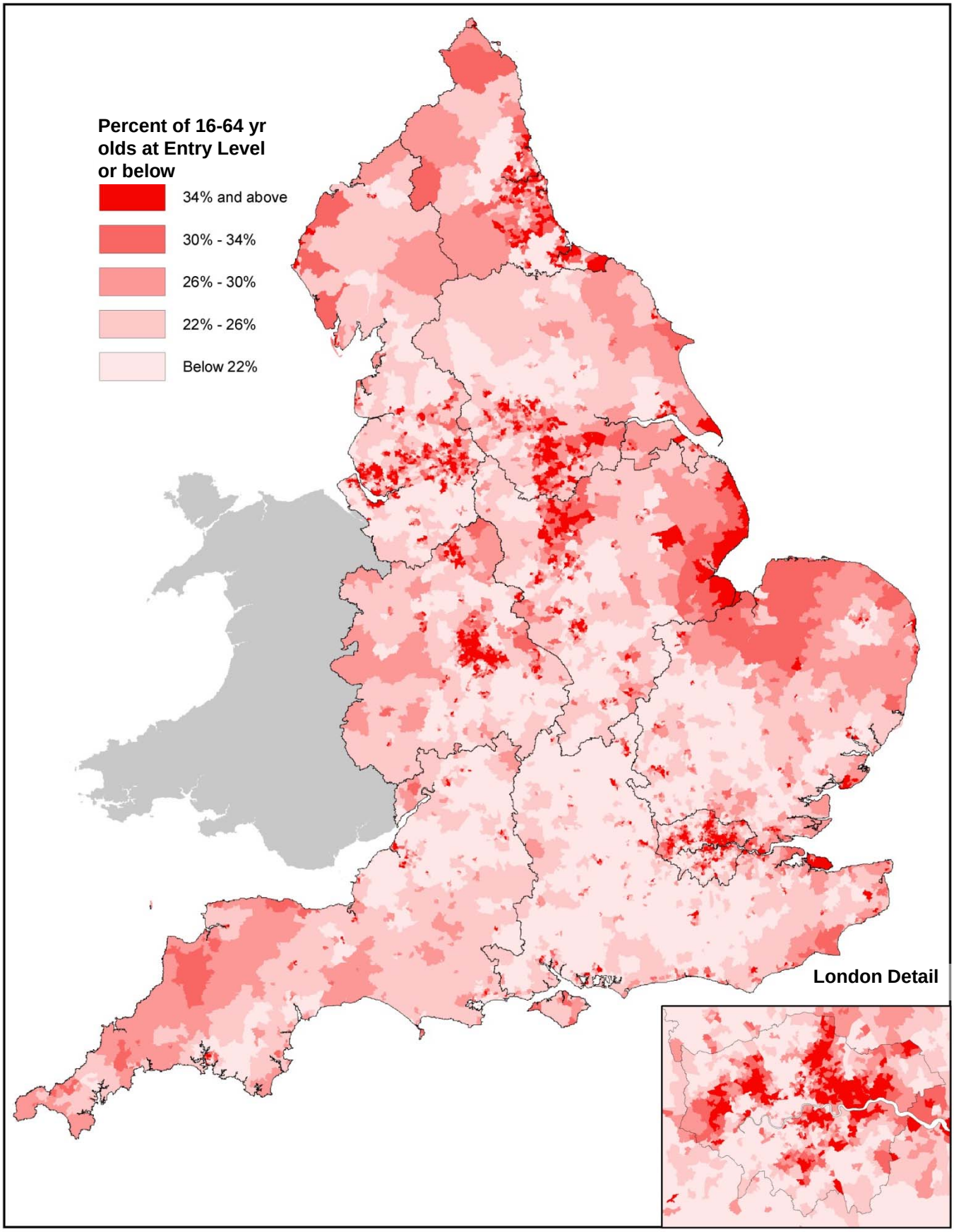


Figure 16 English Spoken as Other Language (ESOL) Estimates: English MSOAs (n=6,781)

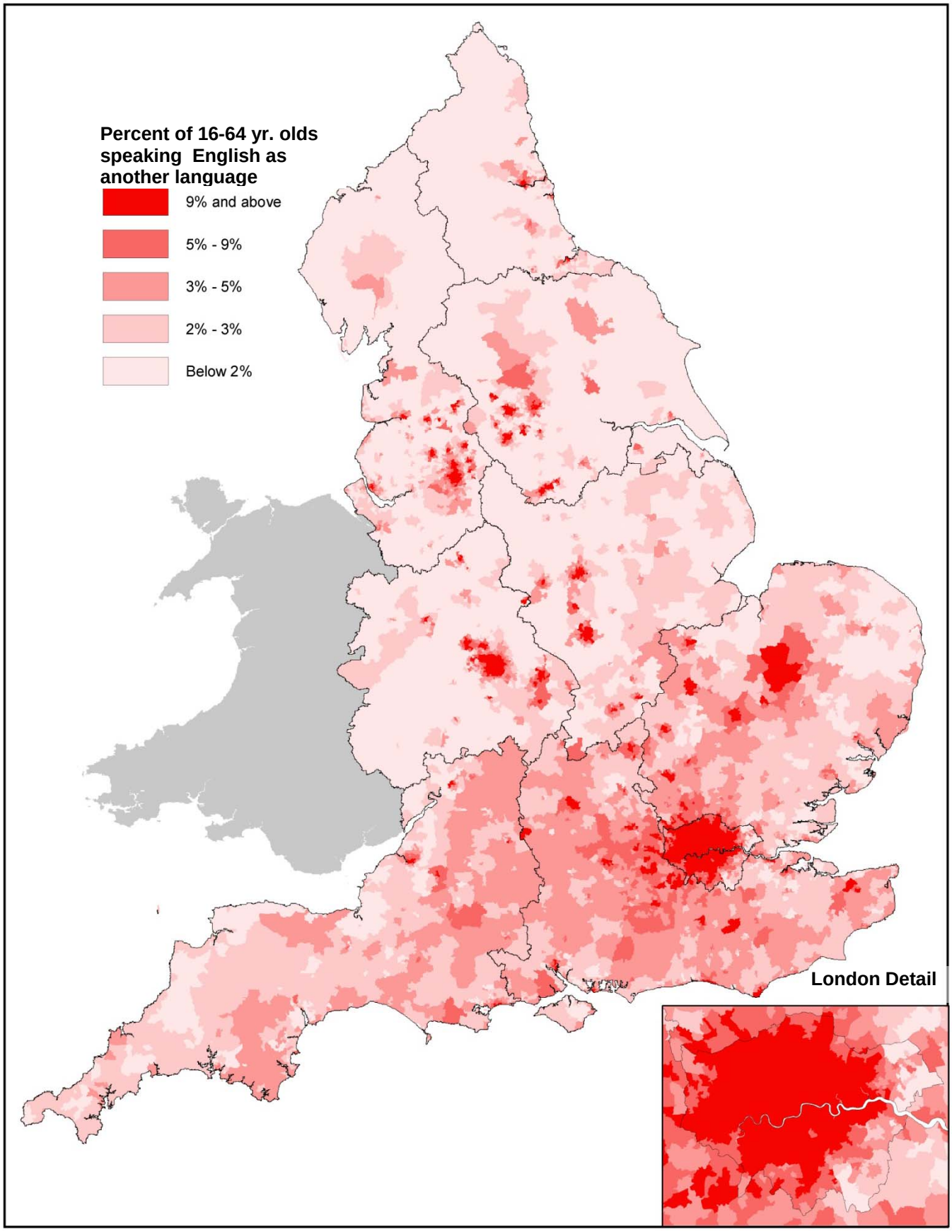


Figure 17 Entry Level and Below Literacy Skills: English Local Authorities (n=326)

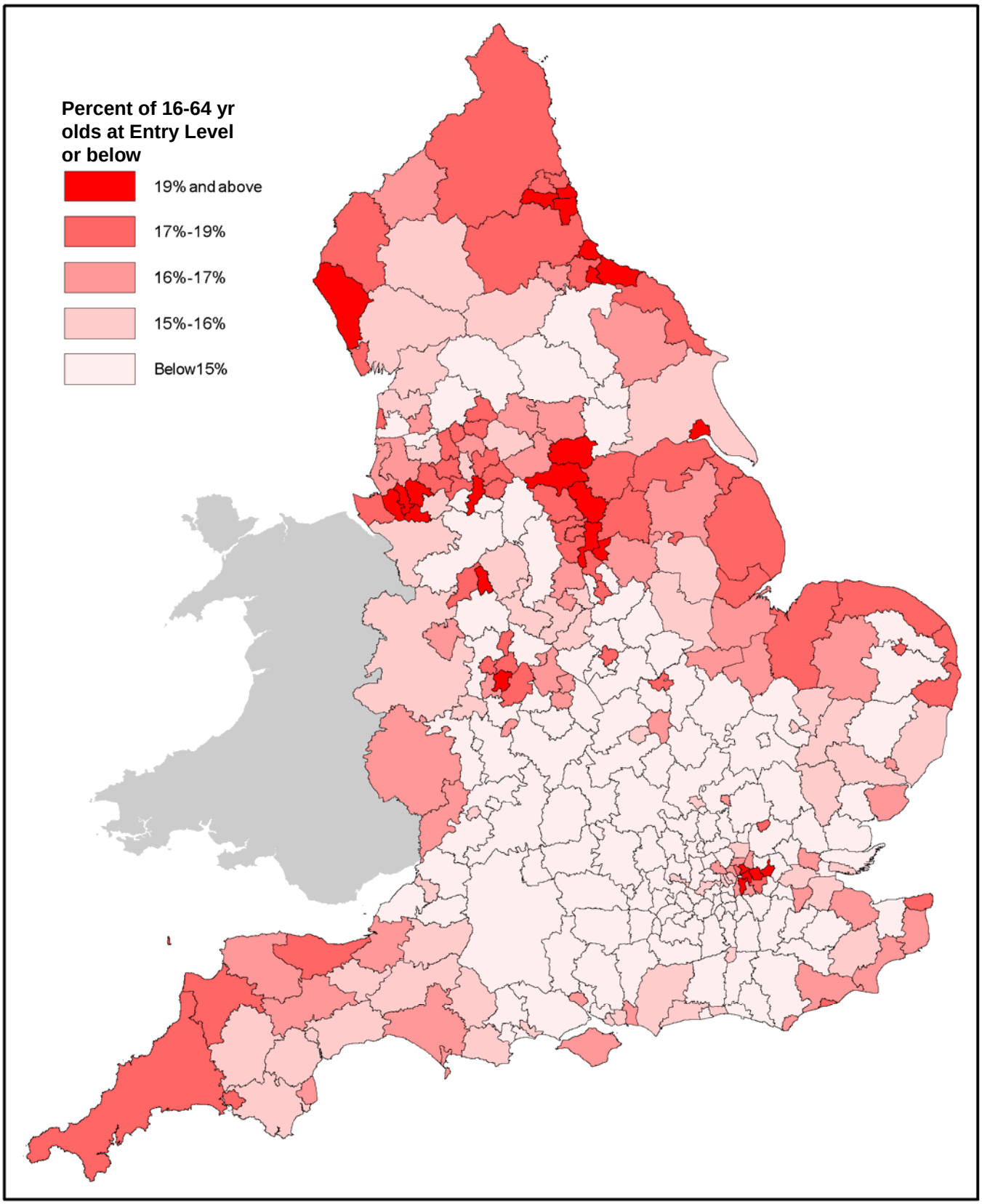


Figure 18 Entry Level and Below Numeracy Skills: English Local Authorities (n=326)

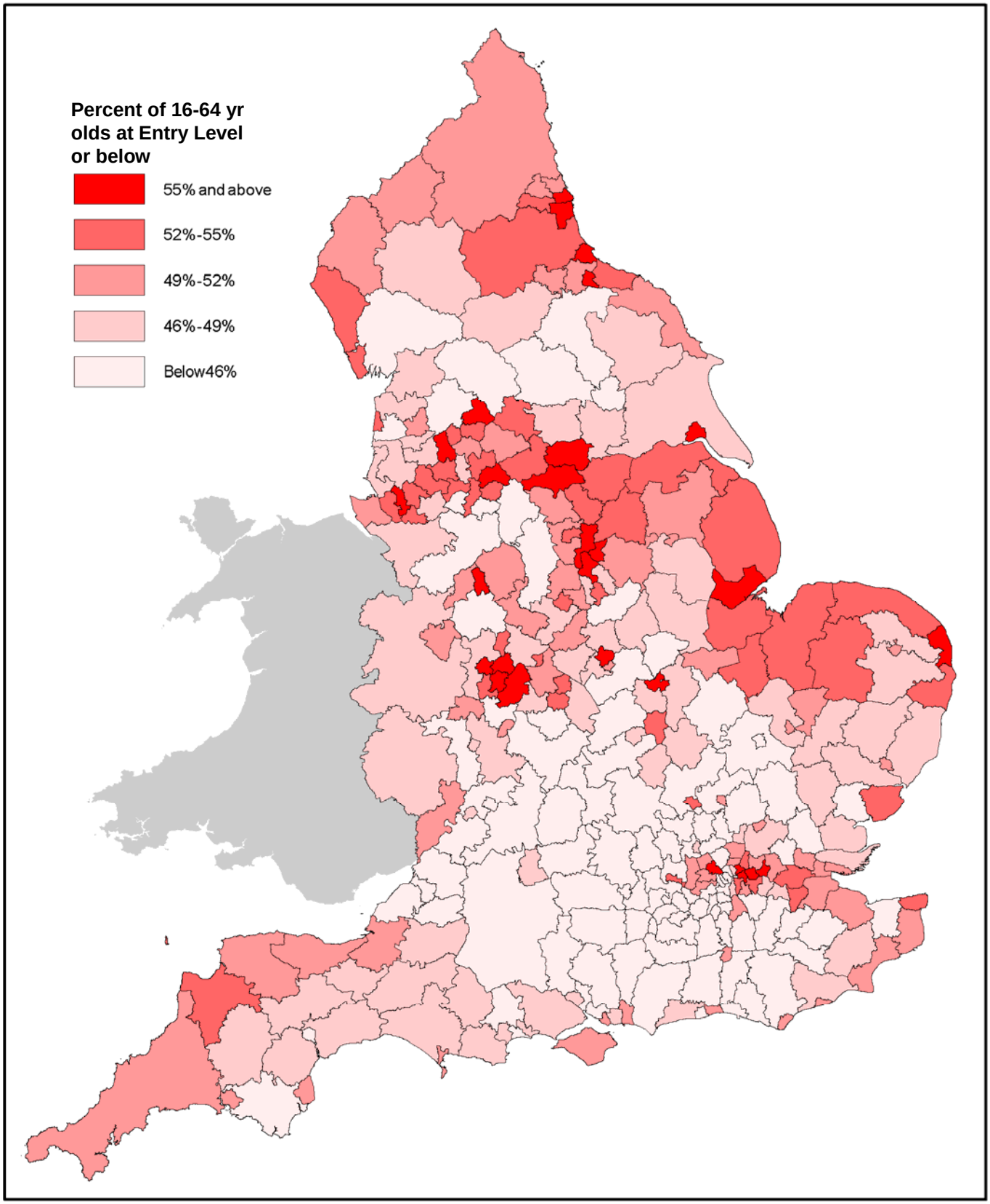


Figure 19 Entry Level and Below Email Skills: English Local Authorities (n=326)

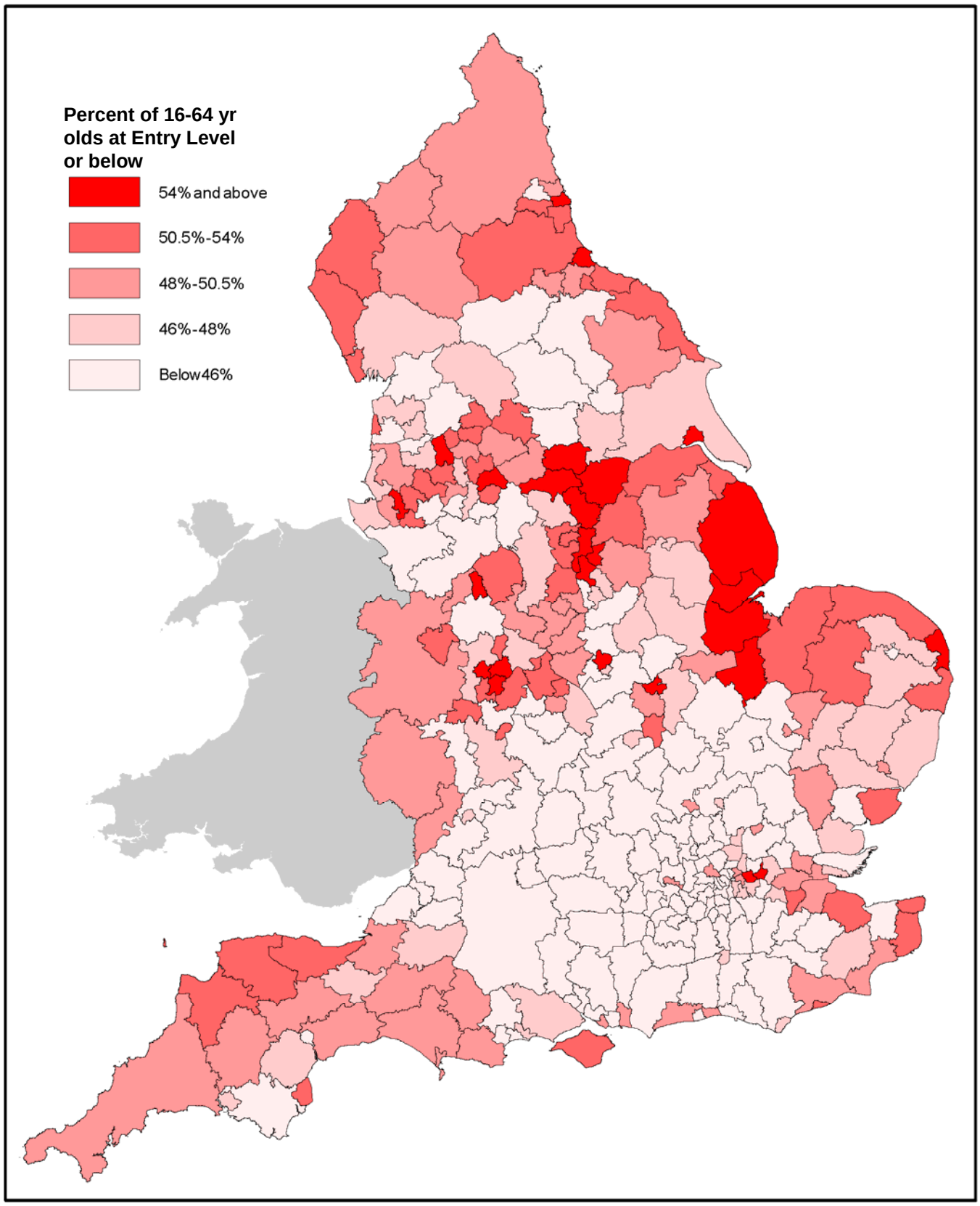


Figure 20 Entry Level and Below Word Processing Skills: English Local Authorities (n=326)

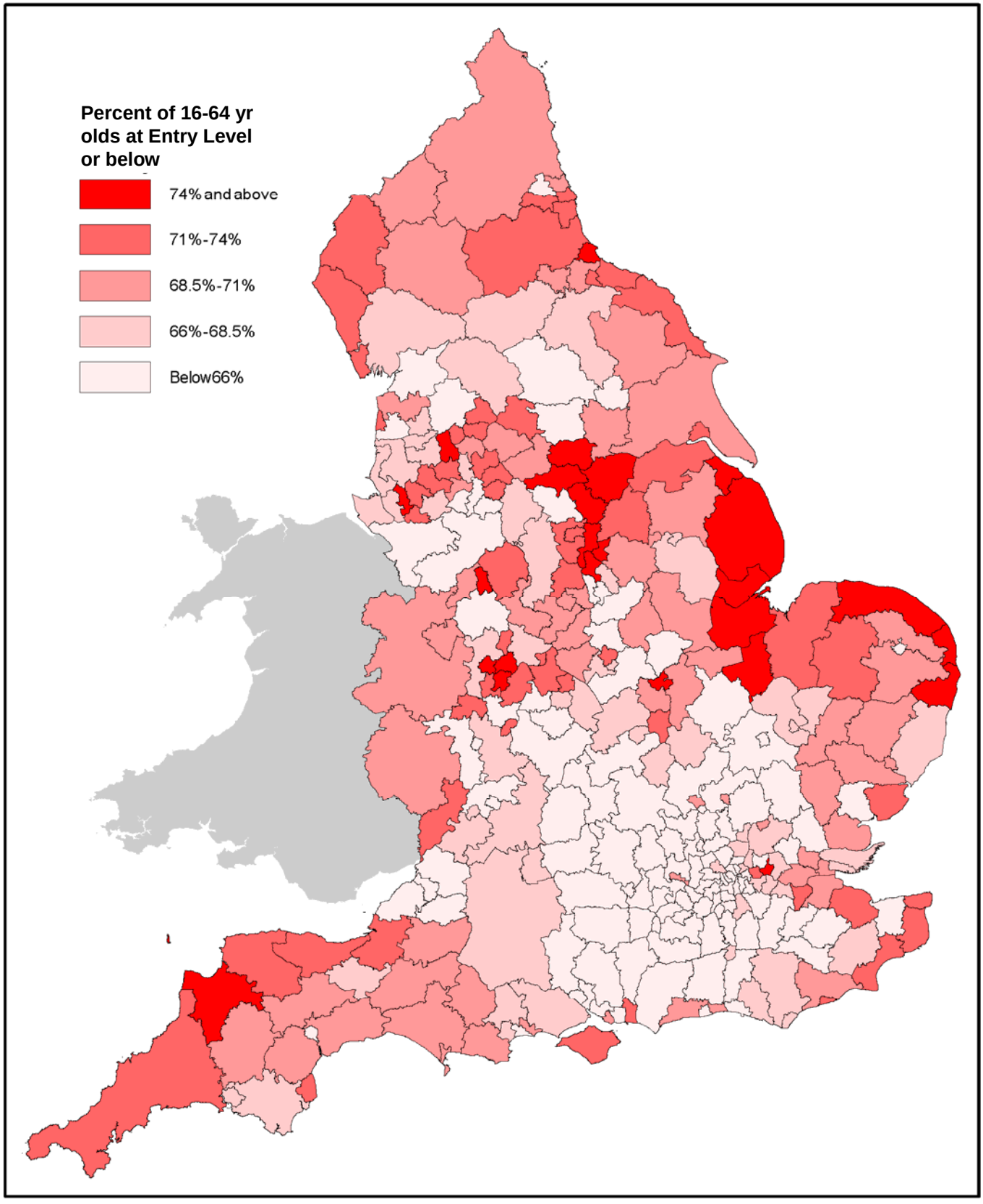


Figure 21 Entry Level and Below Spreadsheet Skills: English Local Authorities (n=326)

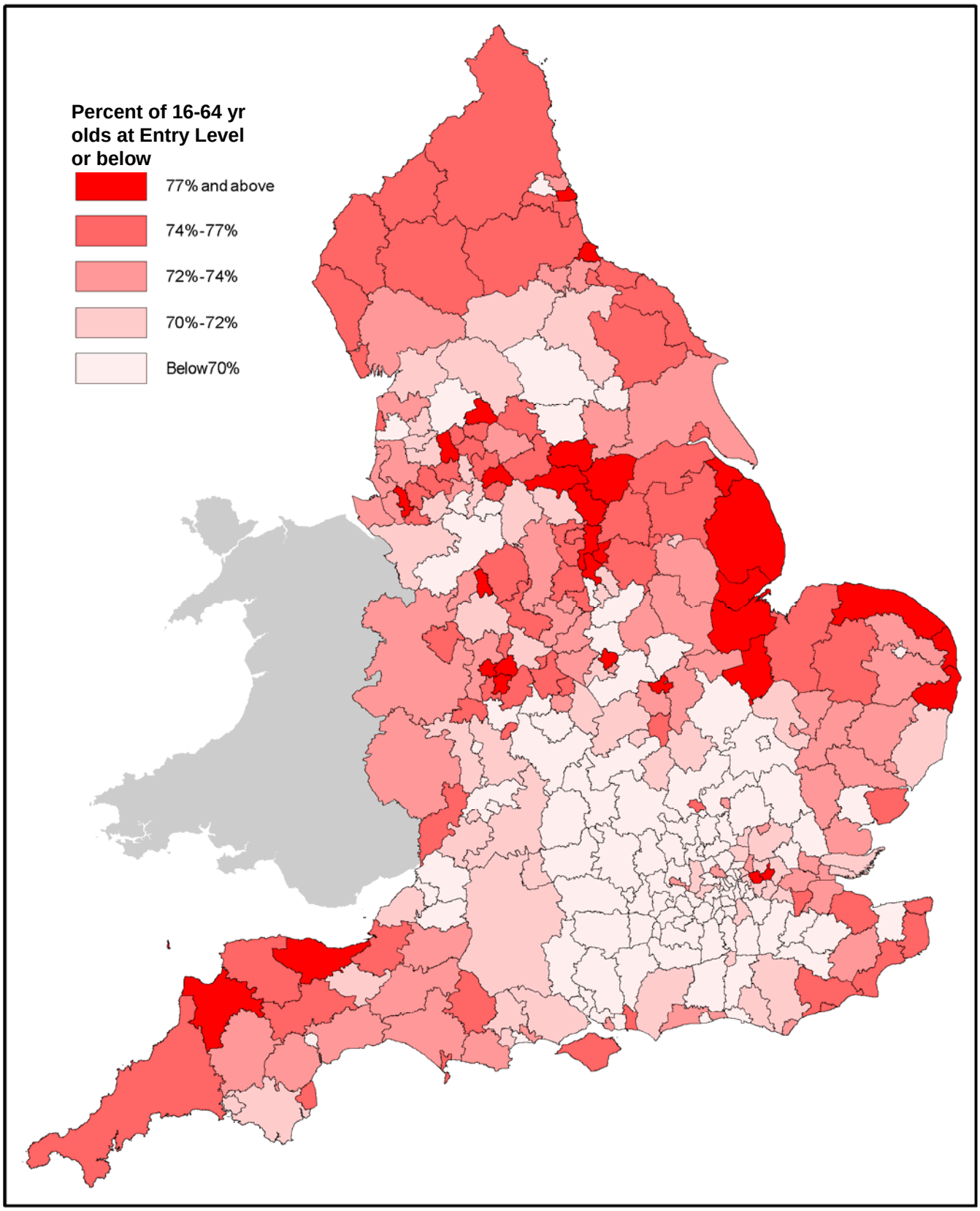


Figure 22 Entry Level and Below ICT (Multiple Choice Test) Skills: English LAs (n=326)

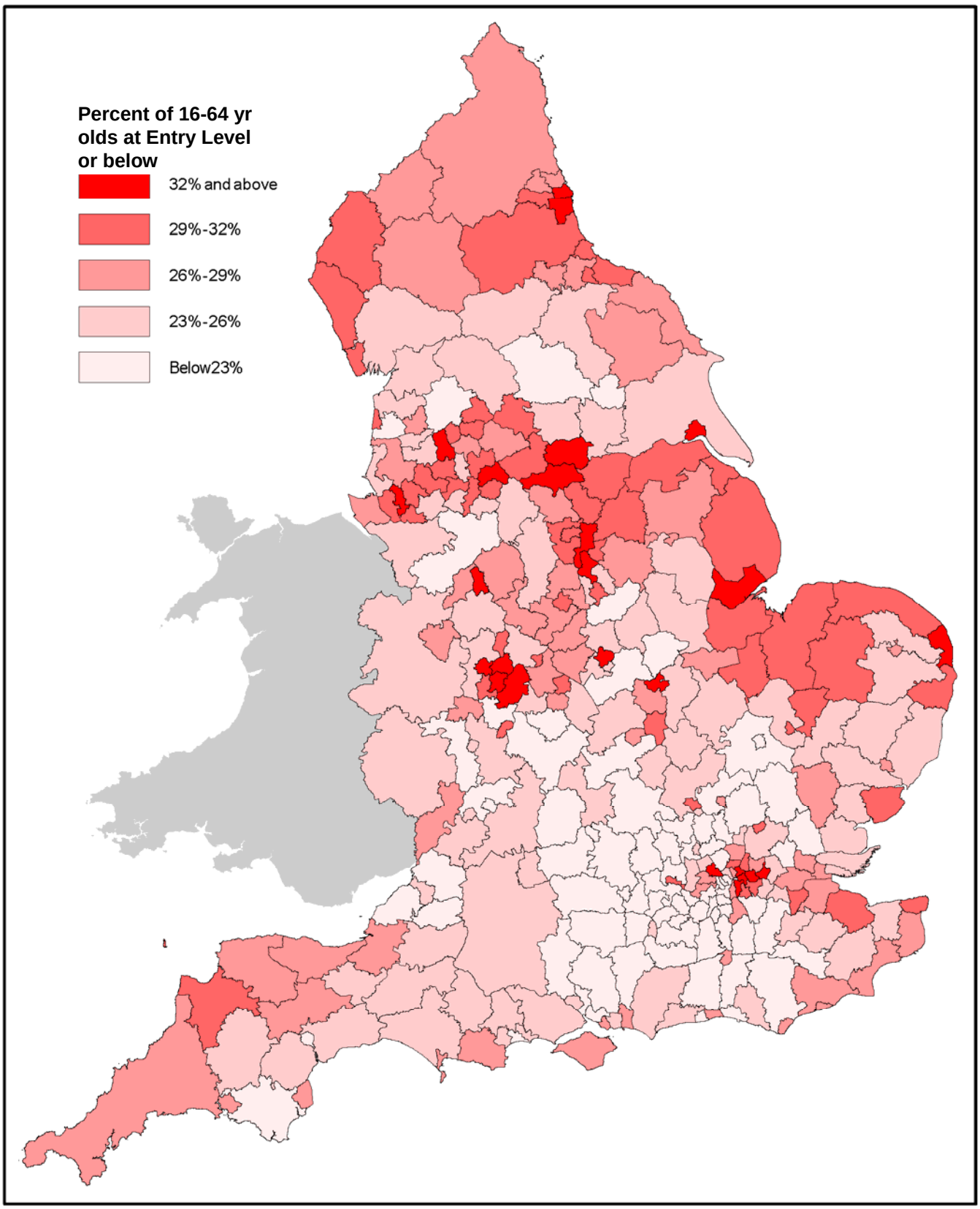


Figure 23 English Spoken as Other Language (ESOL) Estimates: English LAs (n=326)

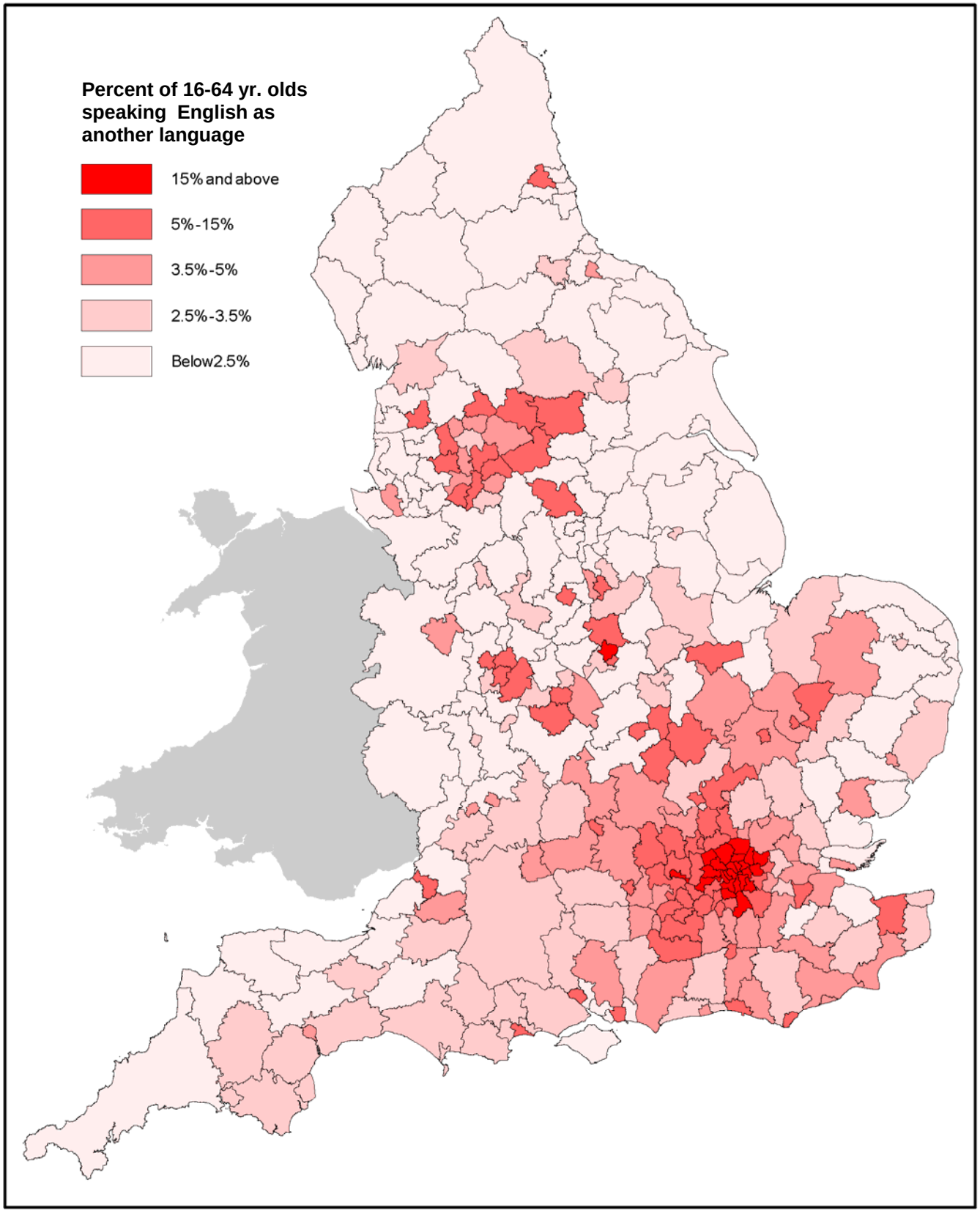
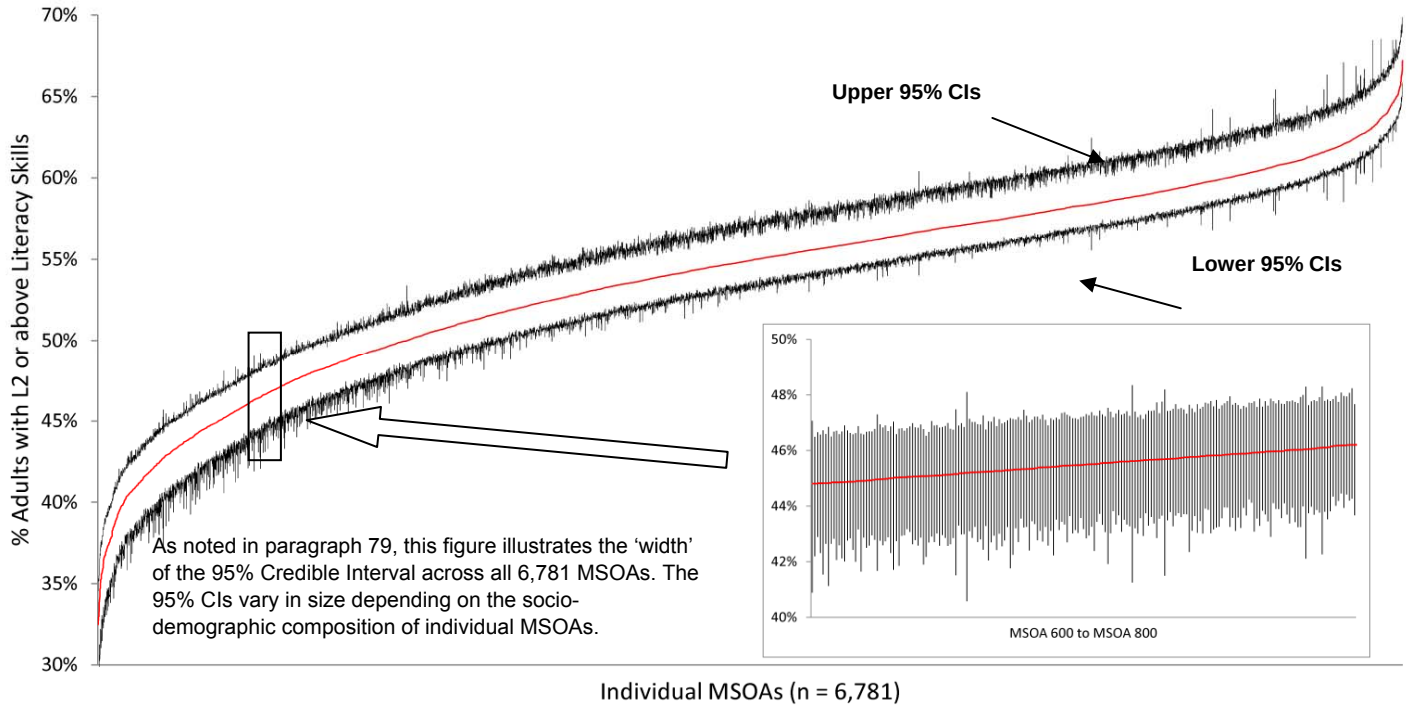


Figure 24 Level 2 and Above Literacy Estimates for MSOAs (2009 Populations): Mean Estimates & Upper and Lower 95% CIs



Summary Guide to Local Area Prediction Excel Files

80. Nine Excel files containing estimates for all reported skills levels have been supplied. All but the first and the last refer to a 2009 population base. These files are as follows:

(a) Middle layer Super Output Areas (MSOAs): 2001 Population Base **(n=6,781)** Estimates have been produced by applying the modelled factor posterior distributions to population covariates derived from the 2001 Census and 4th Quarter 2010 DWP benefit data (weighted to match the 2001 population). These figures have not been adjusted to account for any population change since 2001. ***(middle-layer-super-output-areas-2001-all.xlsx)***

(b) Middle layer Super Output Areas (MSOAs): 2009 Population Base **(n=6,781)** Estimates have been produced by applying the modelled factor posterior distributions to population covariates derived from the 2001 Census and 4th Quarter 2010 DWP benefit data, weighted to fit ONS mid-year age-sex population estimates for 2009. These figures have therefore been adjusted to account for known demographic population changes since 2001, although it has been necessary to assume that MSOA populations have remained static in terms of their socio-economic composition. ***(middle-layer-super-output-areas-2009-all.xlsx)***

(c) Standard Table (ST) Wards: 2009 Population Base (n=7,932) Estimates have been produced by applying the modelled factor posterior distributions to MSOA population covariates derived from the 2001 Census and 4th Quarter 2010 DWP benefit data, weighted to fit ONS mid-year MSOA age-sex population estimates for 2009. These figures have therefore been adjusted to account for known demographic population changes since 2001. These estimates have been proportionally attributed to ST wards on the basis of addresses in the February 2011 *Open National Statistics Postcode Directory* (*ONSPD*). ***(standard-table-wards-2009-all.xlsx)***

(d) 2005 Statistical Wards: 2009 Population Base **(n=7,972)** Estimates have been produced by applying the modelled factor posterior distributions to MSOA population covariates derived from the 2001 Census and 4th Quarter 2010 DWP benefit data, weighted to fit ONS mid-year MSOA age-sex population estimates for 2009. These figures have therefore been adjusted to account for known demographic population changes since 2001. These estimates have been proportionally attributed to 2005 Statistical Wards on the basis of addresses in the February 2011 *ONSPD*. ***(statistical-wards-2009-all.xlsx)***

(e) 2011 Council Wards: 2009 Population Base **(n=7,618)** Estimates have been produced by applying the modelled factor posterior distributions to MSOA population covariates derived from the 2001 Census and 4th Quarter 2010 DWP benefit data, weighted to fit ONS mid-year MSOA age-sex population estimates for 2009. These figures have therefore been adjusted to account for known demographic population changes since 2001. These estimates have been proportionally attributed to 2011 Council wards on the basis of addresses in the February 2011 *ONSPD*. ***(council-wards-2009-all.xlsx)***

(f) **2011 Parliamentary Constituencies*: 2009 Population Base* (n=533)** Estimates have been produced by applying the modelled factor posterior distributions to MSOA population covariates derived from the 2001 Census and 4th Quarter 2010 DWP benefit data, weighted to fit ONS mid-year MSOA age-sex population estimates for 2009. These figures have therefore been adjusted to account for known demographic population changes since 2001. These estimates have been proportionally attributed to 2011 Parliamentary Constituencies on the basis of addresses in the February 2011 *ONSPD*. ***(parliamentary-constituencies-2009-all.xlsx)***

(g) **Local Authorities*: 2009 Population Base* (n=326)** Estimates have been produced by applying the modelled factor posterior distributions to MSOA population covariates derived from the 2001 Census and 4th Quarter 2010 DWP benefit data, weighted to fit ONS mid-year MSOA age-sex population estimates for 2009. These figures have therefore been adjusted to account for known demographic population changes since 2001. These estimates have been proportionally attributed to Local Authorities on the basis of addresses in the February 2011 *ONSPD*. ***(local-authorities-2009-all.xlsx****)*

(h) ***Local Enterprise Partnership Areas: 2009 Population Base*** **(n=37)** Estimates have been produced by applying the modelled factor posterior distributions to MSOA population covariates derived from the 2001 Census and 4th Quarter 2010 DWP benefit data, weighted to fit ONS mid-year MSOA age-sex population estimates for 2009. These figures have therefore been adjusted to account for known demographic population changes since 2001. These estimates have been proportionally attributed to Local Enterprise Partnership Areas on the basis of addresses in the February 2011 *ONSPD*. ***(local-enterprise-partnership-areas-2009-all.xlsx)***

(i) ***Region: Survey Population Base*** **(n=9)** Survey based estimates for Regions have been produced from the survey data. ***(region-survey-estimates-all.xlsx)***

81. Each Excel file contains 14 worksheets. These, as tabulated below, report the number and proportion of adults in each skill level in each geographic unit. Each estimate is accompanied by an upper and lower 95% CI (referring, in other words, to the 2.5% and 97.5% percentiles of each estimate's posterior distribution).

Table 18 Contents of Excel files reporting all skill level estimates

Literacy: (a) counts & (b) proportion	EL1 & below	EL2	EL3	L1	L2 & above	
Numeracy: (a) counts & (b) proportion	EL1 & below	EL2	EL3	L1	L2 & above	
Email: (a) counts & (b) proportion	EL1 & below	EL2	EL3	L1	L2 & above	
Word Processing: (a) counts & (b) proportion	Below EL	EL1	EL2	EL3	L1	L2 & above
Spreadsheets: (a) counts & (b) proportion	EL2 & below	EL3	L1	L2 & above		
ICT Multiple Choice: (a) counts & (b) proportion	Below EL	EL1	EL2	EL3	L1	L2 & above
English not a first language: (a) counts & (b) proportion	Not ESOL	ESOL				

Each 'counts' table also gives the reference population for each local area, which is the number of people aged 16-64 living in households.

82. In addition to these Excel files containing estimates for all reported skills levels, we have also provided a set of files which focus on the number and proportion of adults with either ***Entry Level and below*** or ***Level 1 and above*** skills. The reason for this is that whilst the mean estimates for individual skill levels can be aggregated (i.e. the mean estimate of the number of people with either Level 1 or Level 2 and above literacy skills will be equal to the number of people with Level 1 skills *plus* the number with Level 2 and above skills), the same is not true for the upper and lower Credible Intervals (CIs). Summing these across a combination of categories would result in an unduly wide 95%CI for the aggregated category. The upper and lower CIs should instead be derived *directly* from the set of 1,000 independent estimates of the number of adults in the new combined category.

83. Following discussions with the Department, it was therefore decided to provide direct estimates for Entry Level and below skills and Level 1 and above skills. This, it was suggested, would be the aggregation of individual skill levels that would be of most use to practitioners. A further eight Excel files (covering the same geographies and populations as detailed above) have thus been provided. These carry a "el-l1" suffix, as listed below:

(j) Middle layer Super Output Areas (MSOAs): 2009 Population Base **(n=6,781)** middle-layer-super-output-areas-2009-el-l1.xlsx

(k) Standard Table (ST) Wards: 2009 Population Base (n=7,932) standard-table-wards- 2009-el-l1.xlsx

(l) 2005 Statistical Wards: 2009 Population Base **(n=7,972)** statistical-wards-2009-el-l1.xlsx

(m) ***2011 Council Wards: 2009 Population Base*** **(n=7,618)** council-wards-2009-el-l1.xlsx

(n) **2011 Parliamentary Constituencies*: 2009 Population Base*** **(n=533)** parliamentary-constituencies-2009-el-l1.xlsx

(o) **Local Authorities*: 2009 Population Base*** **(n=326)** local-authorities-2009-el-l1.xlsx

(p) ***Local Enterprise Partnership Areas: 2009 Population Base*** **(n=37)** local-enterprise-partnership-areas-2009-el-l1.xlsx

(q) ***Regions: Survey Population Base*** **(n=9)** region-survey-estimates-el-l1.*xlsx*

84. These tables should satisfy the needs of most practitioners. For users wishing to combine individual skills levels in different ways, determining the mean of the posterior distribution for any new aggregated category is straightforward as the mean of the sum of two variables is the sum of their means, i.e. .

$$\sum(X+Y)=\sum(X)+\sum(Y)$$

Thus the proportion of adults in the new aggregated category will simply be the sum of its constituent categories divided by the total population.

85. Estimating the number of people with each skill level for any aggregation of geographic units is similarly straightforward. For example, two new Local Enterprise Partnerships have been announced since the estimates were produced (Northamptonshire and Buckinghamshire Thames Valley) and estimates for these, or future LEPs or any other units, can be readily calculated by simply adding together the estimates for their constituent LAs. Thus, as illustrated in Table 19 below, Northamptonshire comprises seven local authorities. Estimates (i.e. posterior means) of the number of people with each skill level for each of the constituent LAs are simply added together to provide an estimate of the number of people with those skill levels in the aggregated Northamptonshire LEP.

$$\text{Northamptonshire aggregate mean} = \sum_{i=1}^{7} \text{LAmean}_i$$

86. The percentage of people in with each skill level is then the sum of the means divided by the sum of the populations. Thus, with respect to literacy skills, and drawing the data from local-authorities-2009-all.xlsx;

Table 19 Aggregating Literacy Skill Estimates for Northamptonshire LEP

			Posterior Mean Estimates				
ONS Code	LA Name	Pop	EL1 and below	EL2	EL3	L1	L2 and above
E07000150	Corby	34,715	2,087	959	3,478	11,305	16,886
E07000151	Daventry	49,660	2,076	1,000	3,839	14,547	28,197
E07000152	East Northants.	53,102	2,279	1,094	4,196	15,815	29,717
E07000153	Kettering	57,098	2,497	1,194	4,559	17,037	31,811
E07000154	Northampton	137,281	6,275	2,966	11,166	40,676	76,198
E07000155	South Northants.	56,436	2,135	1,041	4,064	16,101	33,095
E07000156	Wellingborough	47,561	2,360	1,112	4,159	14,639	25,290
Northamptonshire LEP Counts		435,853	19,710	9,366	35,462	130,121	241,196
Northamptonshire LEP %			4.5%	2.1%	8.1%	29.9%	55.3%

87. Dealing with the 95% CIs is less straightforward. In the data provided these have all been derived directly from the posterior estimate distributions, with the upper and lower 95% CIs representing, respectively, the 2.5 and 97.5 percentiles of the 1,000 independent estimates that comprise each posterior distribution. It is possible, however, to algebraically approximate the upper and lower 95%CIs of any aggregated category as long as one assumes that each individual set of posterior estimates is normally distributed. Whilst many of the distributions are somewhat skewed, in practice the assumption seems reasonable insofar as algebraic estimates for the 'Entry Level and below' and 'Level 1 and above' categories are very similar to those obtained directly, albeit that the algebraic approximation appears to slightly underestimate the width of the 95%CI.

88. If one assumes normality, then the upper and lower 95% CIs of the aggregated category will be equal to the aggregate mean +/- 1.96 times the square root of the sum of the variances of its component categories. These variances can be estimated, again assuming normality, as the square root of the difference between the upper and lower 95% CIs, divided by 2 * 1.96.

89. The same approach can, of course, be used to combine upper and lower CI estimates for a set of areas to approximate the upper and lower CIs for a single aggregated area. Thus again using the example of the recently created Northamptonshire LEP, the upper and lower 95% CIs for any particular skill level can be estimated as follows:

$$\text{Northamptonshire LEP upper/lower CIs} = \sum_{i=1}^{7} \text{mean}_i \pm \left(\sqrt{\sum_{i=1}^{7} \left(\frac{(\text{upperCI}_i - \text{lowerCI}_i)}{(2 \times 1.96)} \right)^2} \times 1.96 \right)$$

90. The upper and lower CIs can then be expressed in percentage terms by dividing by the sum of the populations. Table 20 below illustrates how the upper and lower 95% CIs for Northamptonshire LEP can be algebraically estimated on the basis of the

upper and lower CIs of its constituent LAs. Thus, with respect to literacy skills, and drawing data from local-authorities-2009-all.xlsx:

Table 20 Estimating the Upper and Lower 95% CIs for Entry Level 1 & Below Literacy Skills; Northamptonshire LEP

		EL1 and below		Difference (upper - lower)	Divide by (2*1.96) = st. dev.	Squared = variance
ONS Code	LA Name	Lower 95% CI	Upper 95% CI			
E07000150	Corby	1,837	2,373	536.45	136.85	18,727.78
E07000151	Daventry	1,816	2,388	571.60	145.82	21,262.40
E07000152	East Northants.	1,996	2,619	623.34	159.02	25,285.87
E07000153	Kettering	2,182	2,877	694.31	177.12	31,371.46
E07000154	Northampton	5,542	7,132	1,590.20	405.66	164,562.68
E07000155	South Northants.	1,858	2,495	637.21	162.55	26,423.66
E07000156	Wellingborough	2,073	2,703	629.29	160.53	25,770.90
						313,404.75
		Mean – 1.96*sd	Mean + 1.96*sd		Square root gives standard deviation of:	559.83
18,612	20,807					
4.3%	4.8%					

91. The same method can then be used to derive estimates of upper and lower 95% CIs for the remaining literacy skill levels. The process discussed from paragraph 85 et seq. will have to be repeated for the numeracy, ICT and ESOL estimates for the Northamptonshire LEP and, if required, for Buckinghamshire Thames Valley LEP and any additional aggregated geographic units.

Glossary of Terms

2005 Statistical Wards: In 2003 a policy was introduced across National Statistics to minimise the statistical impact of frequent electoral ward boundary changes, particularly in England. Under this policy any changes to English or Welsh ward boundaries laid down in statute by the end of a calendar year were implemented for statistical purposes on 1 April of the following year, irrespective of the year the actual change came into operation. The wards resulting from this policy were known as 'statistical wards'. A change of policy meant that the last set of statistical wards were for 2005, and their composition in terms of postcode addresses and Census Output Areas is detailed in the *Office for National Statistics Postcode Directory (ONSPD)* Open February 2011 edition, 2011. Available at http://www.sharegeo.ac.uk/handle/10672/203. [Accessed 12/1/2012.]

2011 Council Wards: Also known as Electoral Wards/Divisions, these are sub-divisions of Local Authorities and are the key building block of UK administrative geography, being the spatial units used to elect local government councillors in metropolitan and non-metropolitan districts, unitary authorities and the London boroughs in England. The ward geography used in this report is that current as of the 1st January 2011, and is detailed (in terms of postcode addresses and Census Output areas) in the *Office for National Statistics Postcode Directory (ONSPD)* Open February 2011 edition, 2011. (http://www.sharegeo.ac.uk/handle/10672/203.) [Accessed 12/1/2012.] Council wards have a Government Statistical Service code starting E05.

2011 Parliamentary Constituencies: English Parliamentary Constituencies relate to those defined by the Parliamentary Constituencies (England) Order 2007 and the Parliamentary Constituencies (England) (Amendment) Order 2008. They came into effect at the May 2010 General Election. No further changes are envisaged until 2014/2015. Their composition in terms of postcode addresses and Census Output Areas is detailed in the *Office for National Statistics Postcode Directory (ONSPD)* Open February 2011 edition, 2011. Available at http://www.sharegeo.ac.uk/handle/10672/203. [Accessed 12/1/2012.]

Credible Interval: In Bayesian statistics, a credible interval (or Bayesian confidence interval) is an interval in a posterior probability distribution used for interval estimation. A 95% credible interval is thus defined as the range within which 95% of the posterior estimates lie and, on that basis, we can state that we are 95% certain that the true parameter value lies within the stated range.

Generalised Linear Mixed Models (GLMM): In statistics, a generalized linear mixed model (GLMM) is a particular type of mixed model. It is an extension to the generalized linear model in which the linear predictor contains random effects in addition to the usual fixed effects. Fitting such models by maximum likelihood involves integrating over these random effects. In general, these integrals cannot be expressed in analytical form. For this reason, methods involving numerical quadrature or Markov chain Monte Carlo have increased in use as increasing computing power and advances in methods have made them more practical.

Indirect Standardisation: The indirect standardisation approach involves the calculation of prevalence rates for sub-populations within a national survey (defined, for instance, by age and sex), and then applying those rates to equivalent sub-population counts in local

areas. An example of this is provided by Gibson et al. in their analysis of the health needs of local populations.[44] In this, the *Health Survey for England* provides the national survey from which age, sex and social class specific prevalence rates for angina and self-reported 'mental disorder' are derived. These rates are then applied to the corresponding age, sex and social class specific counts of registered persons in each of 539 practices in seven Health Authorities to calculate expected rates of angina and mental health disorders in those practice populations. Indirect standardisation is computationally straightforward but conceptually problematic in that it assumes that the local prevalence of, say, mental disorder, is entirely dependent upon the socio-demographic characteristics of the area. It is assumed, in other words, that there is no contextual effect relating to, for instance, place of residence.

Iterative Proportional Fitting: This technique (IPF) provides a method of combining marginal distributions (and two- and three-way joint distributions) to derive a full joint distribution – i.e. one which describes the number of individuals in a population with each unique combination of characteristics. It is used here to microsimulate the detailed composition of MSOAs on the basis of a series of marginal (and two- and three-way joint) distributions available from census tables (see paragraphs 54 to 60).

Limiting Long Term Illness: The *2011 Skills for Life Survey*, along with the 2001 Census and, indeed, most other large-scale surveys, asks respondents whether they have any long-standing illnesses and of what types (using very broad categories). Respondents are then asked whether any of these illnesses "limit your activities in any way". If a positive answer is returned the respondent is classed as having a limiting long-term illness (LLTI).

Local Authorities: The local authorities to which this study refers are those 'district level' (or 'lower tier') authorities current as of the 1st January 2011, thereby including the structural changes effected on 1st April 2009. There are a total of 326 local authorities; comprising 36 metropolitan districts (E08*), 201 non-metropolitan districts (E07*), 56 unitary authorities (E06*), 32 London boroughs and the Corporation of the City of London (E09*). Their composition in terms of postcode addresses and Census Output Areas is detailed in the *Office for National Statistics Postcode Directory (ONSPD)* Open February 2011 edition, 2011. Available at http://www.sharegeo.ac.uk/handle/10672/203. [Accessed 12/1/2012.]

Local Enterprise Partnership (LEP) areas: When the estimates in this report were undertaken there were 37 LEPs, as listed in the Excel file (*local-enterprise-partnership-areas-2009-all.xlsx* and *local-enterprise-partnership-areas-2009-el-l1.xlsx*). Shortly thereafter, in late September 2011, a 38th LEP – Northamptonshire – was announced. A further LEP has been announced more recently – Buckinghamshire Thames Valley. These two LEPs have not been included in the analysis. Each LEP area comprises a number of local authorities, and a number overlap. The definitions reported in http://www.bis.gov.uk/assets/biscore/economic-development/docs/l/12-p113b-local-authority-areas-covered-by-leps.xls (accessed 22/04/2012) have been used to define the composition of each LEP used in this study. It is likely that additional LEPs will be formed over the coming years, and basic skills estimates for these, as well as for the

[44] Gibson, A., Asthana, S., Brigham, P., Moon, G. and Dicker, J. (2002) "Geographies of Need and the New NHS: Methodological Issues in the Definition and Measurement of the Health Needs of Local Populations". *Health and Place*, 8(1), pp47-60.

Northamptonshire and Buckinghamshire Thames Valley LEPs, will have to be approximated using the method described in paragraph 84 et seq..

Lower Layer Super Output Areas: There are 32,482 LSOAs in England (34,378 in England and Wales). They comprise the lowest level in a stable hierarchy of units devised by the ONS for the collection and publication of small area statistics. See entry for Middle Layer Super Output Areas.

Markov chain Monte Carlo (McMC) simulation techniques: Markov chain Monte Carlo (McMC) methods (which include random walk Monte Carlo methods) are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. See paragraphs 15 to 17.

Microsimulation: In general terms, microsimulation is defined as a modelling technique that focuses on, and/or operates at, the level of individual units such as persons, households, vehicles or firms. Within such models each unit is represented by a record containing a unique identifier and a set of associated attributes – e.g. a list of persons with known age, sex, marital and employment status. The term is used in a number of different contexts, but here we draw on the idea of 'spatial microsimulation' which refers to techniques that allow the characteristics of individuals living in a particular area (i.e. small area microdata) to be approximated, based on a set of 'constraint variables' that are known about the area. (Ballas, D., Dorling, D., Thomas, B., & Rossiter, D. (2005). *Geography matters: simulating the local impacts of national social policies*. Joseph Roundtree Foundation. doi:10.2307/3650139 (Available at http://www.jrf.org.uk/publications/geography-matters-simulating-local-impacts-national-social-policies.) [Accessed 16/1/2012.] (See Iterative Proportional Fitting)

Middle Layer Super Output Areas (MSOAs): MSOAs were devised by the ONS as part of a hierarchy of units specifically designed for the collection and publication of small area statistics. There are 6,781 English MSOAs. They are of broadly consistent size (containing about 7,200 people) and are not subject to boundary changes. They were generated by zone-design software which automatically grouped together 2001 census output areas (OAs) into Lower Layer Super Output Areas (LSOAs), and LSOAs into MSOAs, according to a range of designated size, boundary and 'homogeneity' criteria. They have since become the *de facto* standard geography for which most ONS and other administrative data is published. Their composition in terms of postcode addresses and Census Output Areas is detailed in the *Office for National Statistics Postcode Directory (ONSPD)* Open February 2011 edition, 2011. Available at http://www.sharegeo.ac.uk/handle/10672/203. [Accessed 12/1/2012.] Further discussion can be found on the ONS's Neighbourhood Statistics website: http://www.neighbourhood.statistics.gov.uk/dissemination/Info.do?page=aboutneighbourhood/geography/superoutputareas/soafaq/soa-faq.htm. [Accessed 12/1/2012.]

Multilevel models: So called because of the hierarchical (or multilevel) structure by which the data have been collected and/or within which processes are presumed to operate. Individuals are thus 'nested' within areas and (in this instance) their literacy, numeracy and ICT skills are assumed to be a function of both their individual social-demographic characteristics and aspects of the group of which they are a part (in this instance, their MSOA population). Multilevel models can thus account for both compositional and contextual effects.

Posterior Distribution: A posterior probability distribution is the distribution of an unknown quantity, treated as a random variable, conditional on the evidence obtained from an experiment or survey. As discussed in paragraphs 10 to 14, it can be thought of as a distribution of many simulated possible outcomes given the data being modelled and lies at the heart of the modern application of Bayes Theorem.

Small Area Estimation (SAE): Small area estimation is a generic term describing a range of statistical techniques used to estimate parameters for small sub-populations. The sub-populations are usually included as part of a larger survey but sometimes, as in the present analysis, estimates are made for sub-populations which are not actually sampled. (Individuals in the *Skills for Life Survey* were drawn from only 1,516 of the 6,781 English MSOAs for which estimates have to be produced.) Small area estimation aims to overcome the problem that surveys are designed to provide reliable estimates at national and sometimes regional levels – they are not typically designed to provide estimates at lower geographical levels (for example local authorities, wards and MSOAs). To deal with this problem, as described in Section 0, additional data for these small areas are used in order to obtain modelled estimates.

Standard Table (ST) Wards: Census Area Statistics (CAS) wards were created for 2001 Census outputs. In England and Wales they are identical to the 2003 statistical wards except that 25 of the smallest (sub-threshold) wards were merged into seven receiving wards to avoid the confidentiality risks of releasing data for very small areas. As some Census data would not have been confidential if released for CAS wards, another set of wards, known as **Standard Table (ST) wards**, were introduced. These are also based on the 2003 statistical ward set, but this time a total of 113 wards (those with fewer than 1,000 residents or 400 households) have been merged. See http://tinyurl.com/63w8tct for further details [accessed 12/01/2012]. The detailed composition of CASwards (in terms of postcodes and Census Output Areas) is given in the *Office for National Statistics Postcode Directory (ONSPD)* Open February 2011 edition, 2011 (available at http://www.sharegeo.ac.uk/handle/10672/203 [Accessed 12/1/2012]), whilst a lookup table showing which CASwards were merged into which ST wards can be found at http://tinyurl.com/63w8tct [accessed 12/1/2012].

This publication is also available on our website at www.bis.gov.uk

Any enquiries regarding this publication should be sent to:

Department for Business, Innovation and Skills
1 Victoria Street
London SW1H 0ET
Tel: 020 7215 5000

If you require this publication in an alternative format, email enquiries@bis.gsi.gov.uk, or call 020 7215 5000.

URN 12/1318