Leading education and social research
Institute of Education
University of London

USING SCHOOL CENSUS LANGUAGE DATA TO UNDERSTAND LANGUAGE DISTRIBUTION AND LINKS TO ETHNICITY, SOCIO-ECONOMIC STATUS AND EDUCATIONAL ATTAINMENT

A guide for local authority users

by Michelle vonAhn[1], Ruth Lupton[2], Richard Wiggins[3], John Eversley[4], Antony Sanderson[5] and Les Mayhew[6]

[1] The London Borough of Newham

[2] The London School of Economics and Political Science

[3] The Institute of Education, University of London[+]

[4] London Metropolitan University and ppre limited

[5] Surrey County Council

[6] CASS Business School, City University, London and Mayhew Harper Associates

[+] Address for correspondence:

The Department of Quantitative Social Science, The Institute of Education, 55-59 Gordon Square, London WC1H OAL.

1. Introduction

This guide is the result of a project conducted by Michelle vonAhn of the London Borough of Newham along with colleagues at the Institute of Education, University of London ,and funded by the Economic and Social Research Council's UPTAP (Understanding Population Trends and Processes) programme.

The aim of the project was firstly to map the languages spoken by London school children using newly available data from the Annual School Census (ASC), and then to explore (through a more detailed case study in Newham) how this data might be used in conjunction with other administrative data to understand relationships between language, ethnicity and socio-economic status, and relationships between language and educational attainment, taking into account these other factors.

The results of our statistical analyses are available in a number of publications which we list in Section 7. The maps that we produced appear in a new atlas of London languages, *Language Capital* (Eversley et al. 2010) which is available from the national centre for languages CiLT (www.cilt.org.uk). Some can also be found on-line on the London Education Research Unit website (www.**leru**.org.uk/publications and resources/london maps/index.html *). The maps show in an accessible way the numbers of languages spoken, how many people speak them, and where they are distributed within the capital. We were also able to compare the results with a similar exercise undertaken over a decade ago (Baker and Eversley 2000) and to track changes in the number of speakers of many minority languages over that period.*

This guide is not designed to provide a thorough review of the literature on the links between language, ethnicity, socio-economic circumstances and educational attainment but present our findings to provide practical help for other local authority users who might be thinking of using the ASC language data. Derived from our experience on this project, it covers why local authorities might be interested in language, the nature of the data that is available, how to classify languages for analysis, issues in the analysis and presentation of the data, whether and how language data can be matched to local socio-economic data and some ways in which this might be used. We use examples from London and Newham.

2. Why might local authorities be interested in the languages of school children?

There are many reasons why is important to know about the extent and distribution of languages in the population in general. These include:

- identifying whether and where there are people who speak the languages required for business or enhance service provision for welfare, leisure, tourism, sporting events (CBI 2006, UKCES 2010)

- identifying schools and communities where community languages can be built on as a resource for in-school learning, home-school-community partnerships and community development (QCA 2006, Ofsted 2008)
- understanding requirements for translation or English language support
- understanding cultural, institutional and physical change, since language has implications for the construction of identities, and for the physical and institutional architecture such as the names of places and streets, the languages of education, and the policies and practices of government (Chriost and Thomas 2008)

There are many questions that individual users will want to answer, for example: Is the possession of a different first language an asset or not, at school and in the labour market? Does language diversity help or hinder community cohesion? What benefits or disadvantages come from living in a linguistically diverse area?

Language can also give some indicators of place of origin and ethnicity. Knowing about the distribution of languages can thus provide evidence between decennial Censuses of Population about migration and the development of ethnic communities in particular places.

There are, however, currently no datasets other than the Annual School Census that provide comprehensive language information. Aspinall (2007) notes that in the UK, by comparison with other countries, language data are "conspicuously absent" from the Census of population, most major government surveys and national health datasets. Some information has been collected on fluency in English in surveys on adult literacy, in the Health Survey for England and in the Labour Force Survey, which every three years asks about the first language spoken at home. The specific language spoken, if other than English, Gaelic, Welsh or Ulster Scots is not recorded. A question on language will be included in the 2011 Census. This will ask if the 'main language' is English or other, with a write-in space to specify 'other'. The Census will also ask about the level of fluency in English. However, until then (and possibly after then, given the open-ended nature of the question), the ASC data is only way to get a window on the distribution of languages in the population generally, notwithstanding the complexities of estimating up from the school-age population to the population as a whole.

In addition to its value in relation to language distributions generally, data on the languages spoken by school children is important in its own right, for anyone who is interested in helping children enjoy and achieve in education. It is widely recognised that language is important in education, although views differ about why. Some people argue that what is important is English fluency, and this has tended to be the approach followed by government. Funding through Section 11 of the 1966 Education Act, and since 2000 through the Ethnic Minority Achievement Grant (EMAG), has tended to be directed towards provision of English as an Additional Language (EAL) teachers, bilingual assistants or home-school liaison workers. Recent years have seen new developments, including programmes aimed at advanced bilingual students and a "New Arrivals Excellence Programme" aimed at managing increased migration from European countries, but these have still focused on remedying deficiency in English.

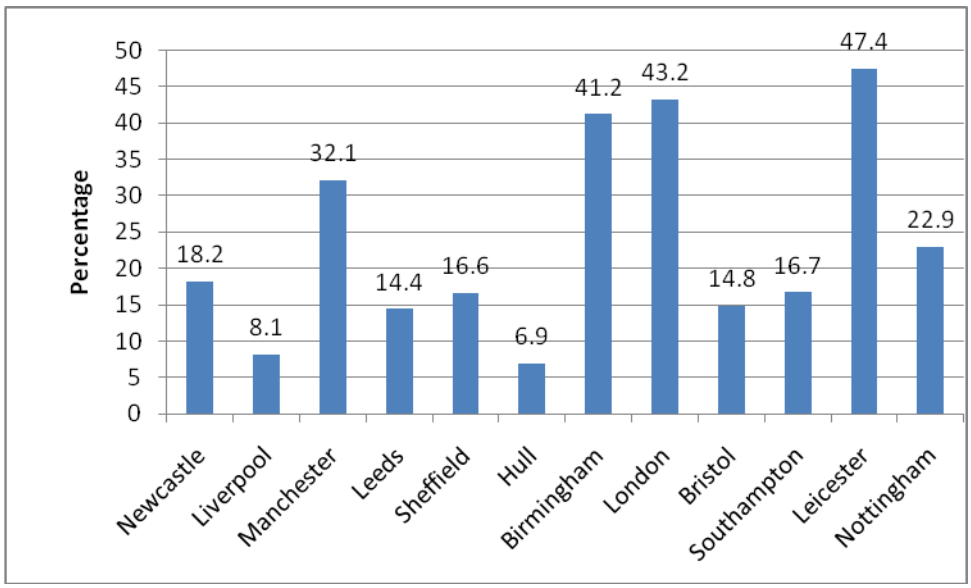
Other people argue that this approach overlooks the benefits of bilingualism to individuals. CiLT (2006, p1) refers to a "substantial" body of research showing that bilingualism is valuable, including that bilingual children perform better than their monolingual peers in tests, especially where they have had the chance to develop both languages in an academic context. The level of development in the first language is a good predictor of development in the second, which suggests that maintenance of first languages is important, quite apart from the value of maintaining languages in which some expertise has already gained, and the probable benefit in adulthood of speaking more than one language. If this is the case, it is important to know not just which children do not speak English, but which languages they do speak, so that appropriate support can be given, including in some cases drawing on resources in the wider local community.

Other researchers have pointed to the relationship between language and socio-economic status. For example Demie and Strand (2006) have shown that EAL speakers actually make more progress between 14 and 16 than English-only speakers once they have become fluent, but this apparent advantage of bilingualism disappears after controlling for other individual and socio-economic factors, suggesting that it is not necessarily the acquisition of English nor competence in another language that affects attainment, but the fact that those who acquire fluent English tend to be from more advantaged backgrounds than those who do not. These relationships are important to untangle. One example we came across was in Surrey, where resources were being allocated to schools partly on the basis of the presence of disadvantaged ethnic groups. One school with a high 'White European' population had low attainment but was not getting extra resource. Further investigation revealed that many of the low attainers were Portuguese speakers whose parents were in low paid jobs and had poor education themselves. Subdividing the 'White European' category by language categories has subsequently led to a change in the basis of resource allocation, to help target sub-groups of this kind.

There is also the question of whether the diversity of languages spoken in a school matters, as well as a student's own language. It might be argued that community languages are easier to support in schools where a number of people speak the same language, and that extreme diversity might be hard to manage. On the other hand, there might be benefits to being exposed to a wide variety of languages. Duckworth et al. (2009) reviewing findings from other literature, conclude that students make less progress in schools with a high proportion of EAL students, although it is not clear what contribution language diversity makes to this, compared with other characteristics of these schools.

Knowing about who speaks which languages, in particular schools and communities, and about how these patterns are changing, is therefore vital in getting a better understanding of the ways in which language matters in education, and what kinds of services are needed where. These issues are becoming increasingly important in some of Britain's major cities. Figure 1 shows that in January 2009, over two-fifths of primary school pupils in London, Birmingham and Leicester and significant minorities in several other cities had a first language other than English.

Figure 1: Percentage of Primary School Pupils Whose First Language is Not English: Major English Cities



Source: DFE: SFR08/2009

Language Data in the Annual School Census

Data on languages spoken in schools was not collected in a consistent form until 2008. Prior to this, some local authorities ran their own data collection exercises, but others did not, and there was no standardisation of the categories used.

In 2007 a 'model' language question was inserted into the Annual School Census (ASC) carried out by state schools across England and Wales each January. However, it was not compulsory to ask the language question and as a result the quality of the data was uneven. In January 2008, the question became compulsory and the data is now collected three times a year in the spring, summer and autumn school censuses. Data is required for all pupils aged 5 and over at 31st August.

Schools are asked to report the pupil's first language. The guidance states that "a first language other than English should be recorded where a child was exposed to the language during early development and continues to be exposed to this language in the home or in the community." In the case of pupils who have been exposed to more than one language, including English, the *language other than English should be recorded, irrespective of the child's proficiency in English. It is important to note therefore, that what is being collected is the first language only. Many pupils who were initially exposed to another first language will also be fluent in English.* In the case of older pupils who are no longer exposed to the first language in the home, nor users of it, school area supposed to consult with the pupil or parent to determine which language should be recorded.

A list of 322 language codes is supplied (attached at Appendix 1). Some of these are variants of other languages, for a person may be classified as speaking Bengali (main category) or Bengali (Sylheti) or Bengali (Chittagong/Noakhali) or Bengali (any other).

There remain some limitations with the language data collection, as follows:

- Schools are not obliged to use the full list of language codes. They may continue to use a shorter list of codes i.e. English, Other than English, Believed to be English or Believed to be other than English. These categories are more likely to be used where schools have very small numbers of non English speakers. However in 2008, 7% of pupil records in London had this insufficiency of language detail and a further 1% were missing or refused. In one authority, Westminster, ambiguous cases totalled nearly 28%. Since data collection practices are likely to vary over time, particular care must be taken with analysis of missing data in studies of change from one year to the next.
- Individual local authorities may also specify that a specific subset of language codes should be used within the authority. Some schools will also use main codes whereas others will use variants (sub-codes). This means that information may be being collected to a different level of detail from schools within a local authority as well as between local authorities. Analysis of the sub-codes or variants is particularly problematic when some schools have used them and others have not.
- As with any data collection, input errors are possible, for example, entering SWA for Swahili rather than SWE for Swedish. Cross checking with ethnic code data can reveal common errors. The 2008 data suggests that Ebira, Guarani and Sundanese in particular have been subject to coding errors, being confused with the following categories respectively : Believed to be English, Gujerati and Sudanese.
- The ASC remains a state school exercise. In some parts of England and Wales and specifically London this is a significant gap. For example in Kensington and Chelsea less than 50% of children are believed to attend local state secondary schools. Some may attend state schools in neighbouring areas. In the country as a whole about 7.5% of children attend private schools. The existence of specialist private schools for speakers of other languages such as the Lycée Français or the German School in South West London may lead to specific gaps in the data but in general the high percentage of children who do attend state schools makes the ASC an invaluable source of data.

3. Legal and Ethical Issues in the Use of the Data

The School Census data is one a large number of so-called “administrative data” sources collected by public authorities for management, planning and monitoring purposes. Local authority researchers will be familiar with the law and ethics governing use of this data, and can seek advice from their legal and data protection teams, so we do not cover this issue in detail here. More guidance and information can be found in Anderson et al. (2006), and from the Information Commissioner’s Office (ICO) (2010a and b).

The 1998 Data Protection Act (OPSI 2010) (Section 33) permits the use of administrative data for research purposes, including statistical and historical purposes, without seeking the explicit consent of the subject, provided that:

> *(a) the data are not processed to support measures or decisions with respect to particular individuals, and*
> *(b) the data are not processed in such a way that substantial damage or substantial distress is, or is likely to be, caused to any data subject.*

The use of language data to identify service needs or to understand social processes in order to inform policy and practice does not pose particular problems under this legislation. However, given that the numbers of speakers of particular languages in particular locations can be very small, particular care does need to be taken to avoid any risk of identification of individuals, particularly when analysis combines language with other variables such as age, gender and ethnicity. The view that we took in this project was that the data has to be used with caution, focussing only on the largest language groupings. Many languages have few speakers particularly when considering a single borough, with over half of the languages recorded in Newham having fewer than 20 speakers. Sub-borough analysis, at LSOA level, becomes even more problematic with disclosive counts, where over 1500 pupils were the sole speaker of a language within their LSOA.[1]
Some local authorities will require officers to have Criminal Records bureau checks in order to access and analyse school census data although at the moment this is not a general legal requirement. Users should check the position in their own authority.

4. Making the Data Manageable and Meaningful: Language Classification

For analysis purposes, the number of language categories needs to be collapsed and there are a number of approaches to this.

The most common approach is to group languages according to their linguistic relationship: common characteristics that are not attributable to borrowing. Within this framework, languages thought to be related by descent from a common ancestor are grouped together. The Linguasphere Register (Dalby 1999, revised in 2007) provides perhaps the most comprehensive classification of the world's languages and dialects along these principles. Languages are assigned to one of five language families, with subdivisions:

- Afro-Asian languages (including Egyptic, Semitic and other subdivisions)
- Indo-European languages (including Indic, Germanic, Slavic and other subdivisions)
- Sino-Indian languages (including Tibetic, Himalayic and other subdivisions)
- Austronesian languages (including Hersperonesic, Transpacific and other subdivisions)
- Transafrican languages (including Bantuic and Benuic and other subdivisions)

[1] DCSF (now DfE) disclosure control restricts the publication of cell values of less than 3 or where values of less than 3 may be derived.

Languages that do not fall into any of these families are allocated to 'geosectors': Africa, Australasia, Eurasia, North America and South America, with subdivisions based either on linguistic relationships or geographical proximity.

The linguasphere is reproduced in *Language Capital*. It is particularly useful for users interested in the geographical spread of language and in the extent of mutual intelligibility. However, it has some limitations as a classification system in the UK context and for socio-economic research. Many of the languages have few if any speakers in the UK. Moreover, location of origin can be more useful to know than linguistic structure, when the classification is being used to understand social, economic or cultural background. For example, Gujerati and Danish are both Indo-European languages; Afrikaans, German and English are all Germanic languages, but we would not necessarily want to group these in the same category for analysis of languages spoken in London. Moving to an entirely geographical classification, however, has its own problems. Some languages, such as Spanish, are widely spoken around the world, so on their own give little information about place of origin. Any grouping is likely to be contestable or contested- for example, countries which have struggled for independence from their neighbours may oppose classifications which group them together.

For our research we classified only the languages spoken by at least one pupil in a London school in 2008, excluding English. We grouped them primarily on their location in the world. Eight 'geozones' are identified, as follows: Asia (South), Asia (East), Asia (West/central), Africa (North), Africa (West), Africa (East/Central/South), European Union and Other Europe. In addition, the classification included a category 'international/transnational' incorporating the major languages of Arabic, French, Portuguese and Spanish which are spoken around the world, as well as 'other' languages. Each of the 'geozones' was then subdivided into the individual languages that were most commonly spoken within London (between 2 and 8 languages per geozone), plus an 'other category'. This produced 51 sub-categories (of which 43 were specified languages and 8 were 'other' categories) as well as the 10 top-level categories.

Figure 2 shows this classification, which is effectively a classification of languages other than English spoken in London. We believe that this could be applied as a basis for analysis in any English region or local authority, since London is by far the most linguistically diverse part of the UK. To the best of our knowledge, the classification is being considered by ONS as the basis for the analysis of language data in the 2011 Census.

5. Using the Classification for Analytic Purposes

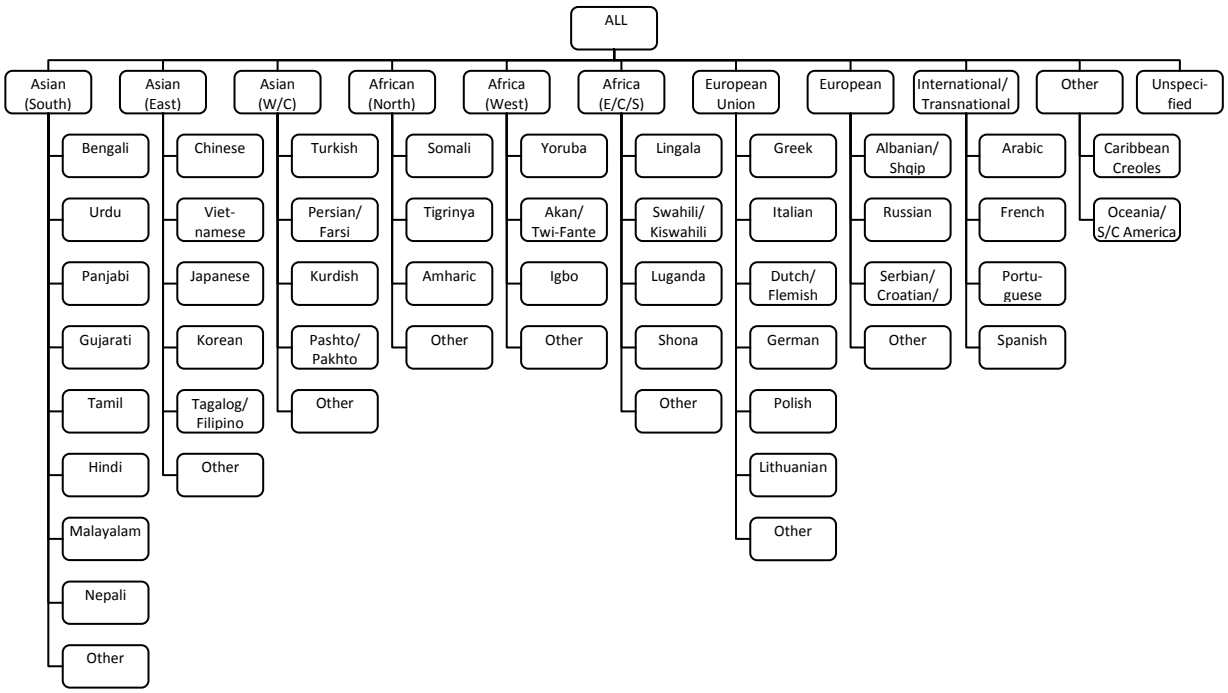
Cell sizes and the Question of Which Categories to Use

A key issue in using the classification at a local authority level is that some individual categories may not contain many if any pupils. Some may therefore become redundant for descriptive purposes. More problematically, there may be too few pupils in some of the

categories to enable robust analysis of the relationships between language and other variables.

Even within London as a whole, with over 1 million pupils of whom over 40% speak a minority language, only 17 of the individual languages were spoken by more than 5000 pupils. Five of these were in the Asian (South) category: Bengali, Urdu, Panjabi, Gujerati and Tamil. Four were the international languages: Arabic, French, Portuguese and Spanish. The others were Chinese, Turkish, Persian, Somali, Yoruba, Akan/Twi-Fante, Polish and Albanian/Shqip.

Figure 2: A Classification of Languages



Users in individual local authorities may therefore want to use a stripped-down classification, removing language categories with very small numbers or collapsing categories. By way of illustration, we show data for all Newham pupils in 2008. There were just over 50,000 pupils, of which 44,140 (88%) had a known language (i.e. not classified as 'other than English', 'believed to be English' and so on). Of these, 97% spoke one of the 43 identified languages within the classification, and 3% were in the 'other' sub-categories. In this borough, the only significant size groupings within these 'other' categories were 'Romanian' (148) and Bulgarian (77), within the 'EU (other)' subcategory. Although small in number, these pupils from EU accession states may represent a minority with particular needs, the existence of which the local authority might wish to note, perhaps extending the classification in this area.

For the 10 main categories (excluding 'unspecified' in this example but adding English), data for Newham were as shown in Table 1. For this borough, each of the main categories is sufficiently populated to provide reasonable sample sizes for analysis. One option would therefore be simply to analyse data by the major categories. However, as Table 1 shows, this would probably provide insufficient elaboration. In Newham, 39% of pupils fall within the 'Asian South' language category, containing speakers of Bengali, Urdu, Panjabi, Gujerati, Tamil, Mayalam and smaller populations speaking Hindi, Nepali and other languages. Bengali, Urdu, Panjabi and Gujerati are all larger in number than most of the main categories, suggesting that, in Newham's case the use of main categories plus individual languages in the Asian South category might be a useful structure for analysis.

Table 1: Numbers in Main Categories (Newham 2008) (excluding unspecified and 'other')

Category	Number of Pupils	Percentage of those classified
English	14710	33%
Asian South	17128	39%
Bengali	*7208*	*16%*
Urdu	*4155*	*9%*
Panjabi	*1652*	*4%*
Gujerati	*1979*	*4%*
Tamil	*1384*	*3%*
Hindi	*222*	*1%*
Malayalam	*506*	*1%*
Nepali	*22*	*0%*
Asian East	814	2%
Asian West and Central	696	2%
African North	1651	4%
Africa West	1982	4%
Africa East Central South	1007	2%
European Union	1468	3%
European Other	496	1%
International	2636	6%
Other	348	1%
TOTAL	42936	97%

The Newham example also illustrates both the value and the problems of imposing categories on the data. Using broad categories provides sufficient numbers in broadly similar groups to enable analysis. However, attention should also be given to the composition of the groups. For example, in Newham, the African (North) category is made up primarily of Somali speakers (1582 out of 1651) and the Asian (East) category is dominated by Tagalog/Filipino speakers, (509) with very few Japanese (7) or Korean(2). This suggests that local authorities need to start with their complete language breakdown and use the classification as a framework to develop the most appropriate structure for analysis in their own particular circumstances. Newham is of course a particularly diverse borough. We welcome comments from users in other areas who apply the classification to their own data.

Creating combined ethnicity/language categories

Another approach which may be fruitful for certain kinds of analysis is to combine ethnicity and language categories. The Census of population (and School Census) uses a five-group classification of ethnicity: White, Black, Asian, Mixed and Other, usually subdivided into sixteen categories as shown in Table 2:

Table 2: Census Ethnicity Classification

Main Code	**Sub-Codes**
White	White British, White Irish, White Other
Black	Black Caribbean, Black African, Black Other
Asian	Indian, Pakistani, Bangladeshi, Asian Other
Mixed	Mixed White /Asian, Mixed White/Black Caribbean, Mixed White, Black African, Other Mixed
Other	Chinese, Other

London data suggests that some of these ethnic groups have a high degree of linguistic homogeneity. For example, in 2008, 84% of pupils identified as Bangladeshi in London spoke Bengali at home (with a further 12% categorized loosely as 'other than English' of which some would be Bengali speakers). 98% of White British and 95% of Black Caribbean children spoke English at home However, other ethnic groups are very linguistically diverse, most notably 'Black African' and 'White Other'.

30% of Black African pupils in London in 2008 spoke English at home, 20% Somali, 9% Yoruba, 6% Akan, 5% French, 2% Lingala, 2% Igbo and 2% Arabic. 179 other languages were spoken by fewer than 2% each of the London's Black African pupils.

Among the 'White Other' ethnic group, Turkish (14%) was the most common language, but 10% spoke Polish, 8% Albanian or Shqip, 6% Portuguese, and 3% each Lithuanian, Greek and Spanish. 'Indian' was also a linguistically diverse category, with two major groups in Gujerati (29%) and Panjabi (23%), as well as Hindi, Urdu, Tamil and Malayalam speakers. For these heterogeneous groups, the collection of data on language provides an opportunity for finer grained understanding of who is living in London and their socio-economic circumstances, and how these are changing over time.

Individual local authorities might therefore find it useful to produce a combined ethnicity/language classification, including the major groups represented. For example, in Newham, the largest ethnic groups among the pupil population in 2008 were Bangladeshi (19%), Black African (19%), White British (13%), Pakistani (13%) and Indian (12%). The vast majority of Bangladeshi pupils spoke Bengali, and the vast majority of White British pupils spoke English. However, the Black African group contained three large linguistic groupings (English, Somali and Yoruba) as well as many smaller groupings that could be classified as indicated above. The largest group among the Indian population was Gujerati speaking, but there were also significant Tamil-speaking and Malayalam speaking groups. Most Pakistani pupils spoke Urdu but there was a significant minority speaking Panjabi. For Newham, therefore, a possible analysis of any socio-economic data might include the main ethnic categories but split the Black African group between English speakers, Somali speakers, speakers of West African languages and others, while splitting the Pakistani group between Urdu, Panjabi and other, and the Indian Group between Gujerati, Tamil, Malayalam and others. Alternatively, language categories could be disaggregated by ethnicity. For example, it might be useful to look at the different socio-economic circumstances and educational achievements of Panjabi speaking children of Indian origin and those of Pakistani origin, or at Gujerati speaking children of Indian origin compared to those of East African origin.

The usefulness of ethnicity/language categories is demonstrated by a preliminary analysis of educational attainment data for London as a whole (for more detail see vonAhn et al 2010). Analysis of attainment at Key Stage 2 in 2008 by ethnic group shows a familiar pattern of differences between ethnic groups. Pupils of Chinese ethnicity were on average the highest attainers. Black Caribbean, Black Other and Black African pupils were the lowest attainers, although within each group there was significant spread of scores and there were high attainers in each group. Analysis by language category begins to illuminate the spread of attainment within broad ethnic categories. For example, with the 'White Other' category, five groups had particularly low attainment. Median scores for Turkish, Portuguese, Lithuanian and Polish speakers would put them at the bottom of the overall distribution.

By contrast, Italian, Greek and English speakers in the 'White Other ethnic category had few low attainers and median scores that place them close to the top of the overall distribution. The Black African category also contained a wide spread. Lingala, French and Somali speakers tended to have very low attainment, well below that of the lowest attaining ethnic group overall (Black Caribbean). However, the attainment of Black African Igbo speakers was similar to that of White British students. These data suggest that some of the commonly used ethnic groupings may be too broad to be useful, and that language data can provide greater insight into which pupils may be in need of particular support. Of course, over time, combining ethnicity and language data will also enable us to identify ethnic communities in which the proportion of minority language speakers is changing (in other words about the differential take-up of English as a home language and the possible loss of bilingualism).

6. Matching ASC language data to other datasets

A key aim of our project was to explore the practicality and value of matching the ASC language data to other administrative datasets in order to estimate the numbers of people speaking different languages in the population as a whole, as well as just in schools, and to gain a fuller understanding of the relationship between language and other socio-economic characteristics. Matching to other data could also be useful in analysis of the relationship between language and educational attainment. The ASC schools census contains information on whether pupils are in receipt of Free School Meals (FSM) but no other information on socio economic status or household characteristics.

Prior to our project, the London Borough of Newham had already commissioned the creation of a matched set of administrative data by consultants Mayhew Harper Associates.

Their approach built on two key datasets: the Local Land and Property Gazetteer (LLPG) from the Local Authority and the patient register data from the Primary Care Trust (PCT). The LLPG is an address register of all properties in the borough, and includes residential and non-residential addresses. The patient register includes names, ages and addresses of people registered with GP practices. The patient register is also a very important source of data with regard to international migration, as new migrants registering with a GP for the first time have their records flagged (commonly known as "Flag 4"). On a first onward move that involves re-registering with a GP, the Flag 4 is lost, and they are simply identified by the NHS Central Register as an internal migrant. In areas of high turnover, the information from Flag 4 is a key source of information about international migrants, and the length of time they appear to stay in the borough.

These two datasets were matched using the address field. Combining these two sources provided a database of individuals within households at addresses within the local authority. Due to poor quality addressing and lag of updating addresses, there will always be a number of patient records that cannot be matched due to lack of full details of address, use of PO Boxes, and changes in the dwelling stock that are not reflected in addressing (splits of properties into flats, primarily). In addition, GP Registers have historically had significant problems with List Inflation or 'ghosts' where people who have died or moved away (or in some cases never existed) have been included. Matching of other datasets via shared identifiers with the LLPG can enable more accurate estimation of the actual population.

In this case other datasets were matched as follows:

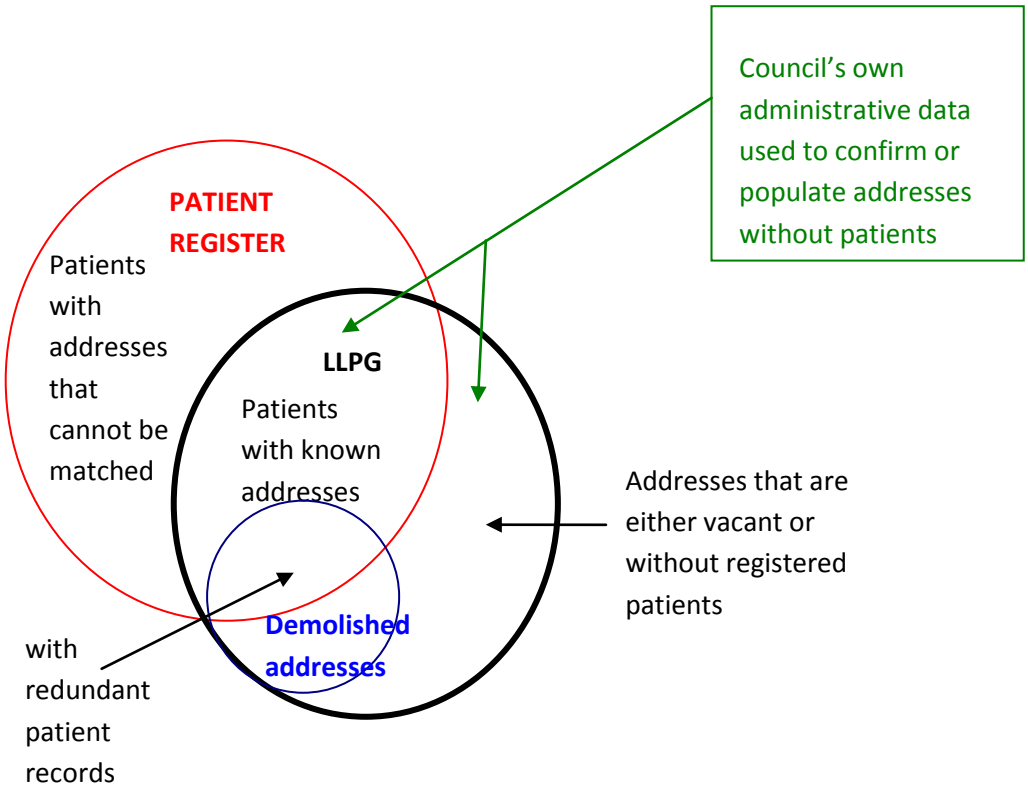
- The Council Tax register. This is a primary "confirmation" data source. The name of the responsible person(s) for Council Tax can validate the occupants in many cases, but there will be addresses within the patient register which are non-residential in nature and excluded from Council Tax. These are communal establishments of various sorts (care homes, hostels, hotels, etc) that will appear in the Non-Domestic Rates register of businesses, and will also include buildings that have been changed without planning permission (shops and offices that are used for residential purposes, or conversions of single dwellings into multiple dwellings). Information on 'vacant properties', single person discount and student discount can also be used as confirmatory sources. Information about whether a household is in receipt of Council Tax Benefit (CTB) provides a proxy for low income.

- The Electoral Register. This is another confirmatory source of data, and is particularly important for confirming multiple names at an address (shared dwellings), and assisting with household structure. As more country of origin data becomes available on the Electoral Register, this will become another key data source for understanding international migration to a borough. At present, it is a useful source for identifying the distribution of European Union voters, but with the fuller collection of nationality of all electors, this could provide useful insight into population diversity and international inflows. There is also scope for obtaining limited age information about some voters (those rising to voting age or those over 70), but this was not used for the project.
- Housing Benefit Data, again for confirmatory purposes

In hindsight, it is recommended that a listing of demolished addresses also be obtained, covering as long a time period as possible, as this can be used to clean the patient register of old and defunct addresses and the attached patients. As there is no requirement to de-register on moving, at least some of the unmatched addresses arising out of the initial linkage to the LLPG can be explained by demolition.

Figure 3 summarises the matching process.

Figure 3: Creating A Matched Dataset



A range of other additional datasets can be used, including library users, school registrations, Blue Badge holders, parking permits, customer contacts, depending how the quality of the data and the intelligence about population characteristics that may be sought. The combined data provides a rich source of intelligence about the population that can help monitor change as well as profiling population segments.

The Annual School Census data is another key data source for confirmation of adult names, with pupil surnames being able to be used to confirm adult surnames that do not appear in any other administrative data.

For this project, in addition to the basic confirmatory data used for the population estimate, we added the ASC language, ethnicity and attainment data, using the Pupil Reference Number to link these variables to the confirmatory data.

Issues in Matching Administrative Data

The creation of such datasets is a reflection of a wider trend across councils which are increasingly linking data across administrative systems in order to facilitate a more joined up approach to managing and delivering services.

Ethical concerns around such approaches relate to:

- Data being used for purposes other than that for which it was collected
- The increased risk of identification when multiple characteristics are combined
- Whether such comprehensive data gathering can be justified if it does not lead to particular social interventions.

Set against this are arguments for:

- *Evidence-based public policy*. There is a general argument that the need for and effectiveness of public policy should be based on the best available evidence.
- *The need for up-to-date comprehensive data*. Policies based on data collected nine years ago in the case of the Census or limited samples as many social surveys are cannot be used to measure the impact of recent policies or social change
- *Use of the cheapest and least intrusive methods.* Collecting new data through surveys when the data is already there is undesirable. The state (in all its forms) holds a great deal of data about people living in the UK which people have a right expect to be used in their best interests
- *Freedom of Information* – the argument that aggregated information about populations and public services should be available to the public (Brooke 2008).

Specifically in this case it may be argued that up-to-date and comprehensive pupil language data, matched to other data sources, can inform debates about how to raise school attainment levels, promote economic competitiveness, plan accessible health services and so on and that absence of such evidence of such data fuels myths that can be not only

harmful to society in general but specifically to school pupils who have to deal with the consequences of ignorance.

If local authorities decide to go ahead with such data matching exercises, they can expect extensive negotiation for access, sharing and use of the data. Use of the patient register data requires a data sharing agreement between the relevant PCT and local authority, covering the precise information being sought, the means by which it will be provided (secure transfer or collected media), how it will be used, and its disposal after use. Use of an intermediary to act as a "safe haven" for all the data and to undertake the matching can help as this ensures that the personal information regarding names is never shared. In this case, the data sharing agreements are between the data owner and the intermediary.

Some datasets in particular are difficult to obtain in full. The electoral register exists in two versions, one full and one edited to take out people who have opted out of the full register. The full register is of greatest use for this work, but can be difficult to obtain, depending on how the local custodian defines what constitutes a legitimate use by the local authority. Generally, if the work will be used to create aggregate information and is not used for any kind of targeted, marketing purposes, it is more likely to be seen as an acceptable use. Analysis of the demography of those not registered to vote can also aid in obtaining approval, as this is directly relevant to the owner of the data. The Annual School Census may also be problematic, if analysis is needed covering more than one borough. The data held by individual local authorities covers only children attending state schools in that borough, so resident children attending state schools in other areas are not included in this data set. The detailed name and address data is not readily available from the Department for Education, but may be made available upon presentation of a significantly strong business case. It is becoming increasingly difficult to obtain such data other than for the purposes of educational analysis (for example studies of residential mobility may not be granted access to the data). This is a key area that needs to be resolved for this work to have complete coverage, either with a multitude of data sharing agreements with other local authorities to share the relevant data for resident pupils or to secure agreement from DfE for the data to be provided for use in this way by local authorities.

Analysis of the Matched Data

Finally, in this section, we present an overview of the kind of analysis that was possible with the matched data. We undertook two analytic exercises:

- Descriptive statistics
- A 'risk ladder' analysis

For the first analysis we took the decision to confine the analysis to pupils of the same age, in order to consider the relationship with educational attainment. Standard assessment data was available in the dataset for Key Stage 2 (age 11) and Key Stage 4 (age 15). The effect of including the test data in the analysis was that we could only consider pupils in these school years, which substantially reduced the sample for analysis. For example in

Newham, the sample of students with known ethnicity at Key Stage 2 in 2008 was just under 3000. The largest group was Black African (547) closely followed by Bangladeshi (537). Other groups were as follows: White British 375, Pakistani 348, Indian 292, Black Caribbean 193, Other Asian 167, White Other 137, Other 112, Any other mixed 82, Black Other 55, Mixed White and Black Caribbean 55, Mixed White and Black Caribbean 31, Mixed White and Asian 15, White Irish 9 and Chinese 9.

We looked at language breakdowns for the two largest groups. 80% of the Bangladeshi pupils spoke Bengali, with another 19% being classified as "other than English". In this case, an analysis of language within ethnic group is not helpful. Within the Black African group, about a sixth spoke English. The largest group of non-English speakers (21% of all Black African pupils) was coded as 'other than English' or 'unknown'. This left only three languages with more than circa 50 pupils at Key Stage 2: Somali, Yoruba and Akan. Some clear differences could be observed between the groups, with Somali speakers being much more likely to be eligible for Free School Meals (FSM) or Council Tax Benefit (CTB) and to be from single parent families. However, this is really too small a cell size for robust analysis and in particular for any statistical analysis of the relationship with attainment.

Table 3: Socio-economic characteristics within the Black African Group in Newham, KS2 2008

	Pupils	3 or more children	Single parent	FSM	CTB	Not SEN
English/Believed to be English	85	58%	21%	32%	38%	73%
Somali	88	89%	41%	91%	97%	69%
Yoruba	65	72%	15%	25%	26%	68%
Akan	47	68%	19%	30%	36%	72%
Other than English/unknown	118	80%	32%	59%	63%	71%

Risk ladder analysis is a method by which clusters of 'risk factors' are identified within individual cases, and then the relationship between the clusters and an outcome examined. This technique clearly requires reasonable sample sizes to show statistically significant results from the inclusion of new variables in a cluster. For illustrative purposes we show this using another 'outcome': the likelihood of being classified as having Special Educational Needs (SEN). This analysis is purely designed to demonstrate the method – we recognise that SEN classification is not an outcome variable per se, incorporates a wide variety of different difficulties and conditions, and is also affected by the propensity of schools to classify pupils (Crawford and Vignoles 2010). The likelihood of being identified with SEN also increases with age. We use it to illustrate that even when an outcome is used for

which all pupils in an educational phase (primary or secondary) can be included, there are still limitations on analysis at local authority level.

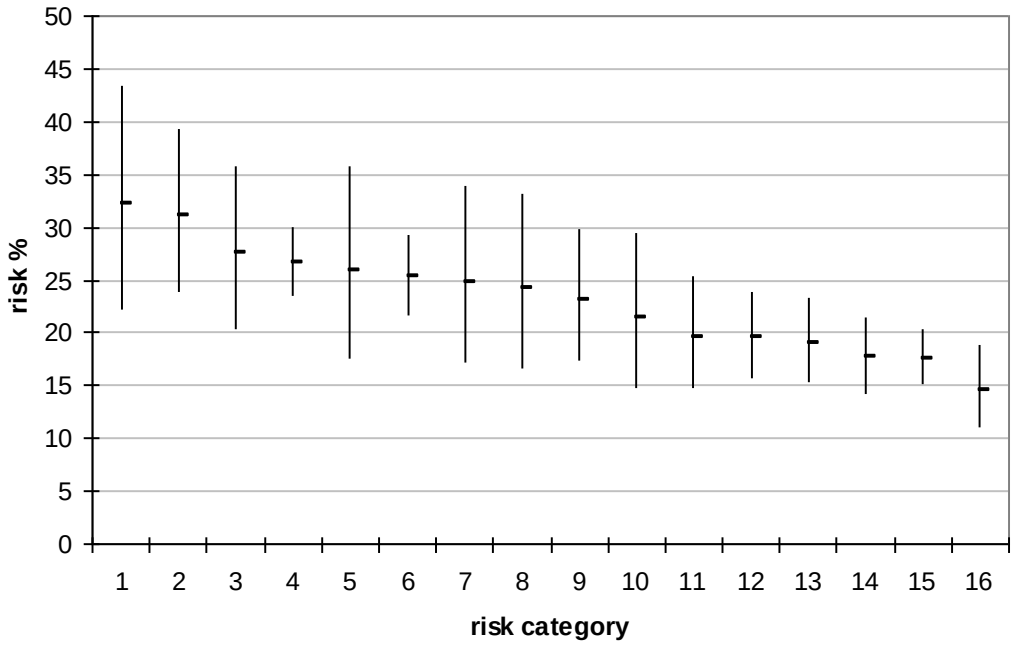
Table 4 shows the clusters within the Newham Black African primary school population (over 3000 children) using four 'risk factors' identified within the matched dataset: whether on Free School Meals, from a single parent family, from a household with 3 or more children and speaking a home language other than English. Finer language classifications could not be included for sample size reasons.

Clusters are listed in order of descending risk of SEN. Cluster 1 includes pupils with FSM but no other 'risk factors', cluster 2 with FSM and a first language other than English and so on. The table demonstrates the influence of poverty. Black African pupils with Free School Meals as a 'risk factor' were more likely to be identified with SEN than those without FSM and in some cases the differences were within the bounds of confidence intervals given the sample size (see the last two columns). However, sample sizes did not produce statistically significant results for comparison of more complex clusters. The confidence intervals (CIs - at 95% level) for some groups are very wide and most clusters overlap. For example, there is no statistically significant difference between cluster 1 (FSM only), cluster 2 (FSM+not English), cluster 10, (FSM+ not English+single parent) and Cluster 6 (FSM+ not English+single parent+ 3 or more children). The differences between the clusters and also the overlapping confidence intervals are shown graphically in Figure 4.

Table 4: Example of Risk Ladder Analysis for Newham Black African primary pupils

Cluster	Frequency	FSM	English not first language	single parent household	3+ children	SEN	lower CI	upper CI
1	81	Y				32.1	22.2	43.4
2	148	Y	Y			31.1	23.7	39.2
3	142	Y			Y	27.5	20.3	35.6
4	746	Y	Y		Y	26.5	23.4	29.9
5	97	Y		Y		25.8	17.4	35.7
6	516	Y	Y	Y	Y	25.2	21.5	29.2
7	113			Y	Y	24.8	17.1	33.8
8	112	Y		Y	Y	24.1	16.5	33.1
9	187		Y	Y		23.0	17.2	29.7
10	131	Y	Y	Y		21.4	14.7	29.4
11	225			Y		19.6	14.6	25.3
12	380				Y	19.5	15.6	23.8
13	400					19.0	15.3	23.2
14	454		Y			17.6	14.2	21.4
15	895		Y		Y	17.5	15.1	20.2
16	324		Y	Y	Y	14.5	10.9	18.8

Figure 4: Graphical Representation of Risk Estimates for Cluster Groups



Overall, our conclusion from this exercise was that while the matched data are potentially very rich and show some interesting patterns in initial analysis, at an individual London borough level, the cell sizes are too small to enable meaningful analysis.

One option for individual local authorities considering a similar analysis is to pool data in order to generate larger samples for analysis. Pooling could be undertaken over several years or with neighbouring authorities. However in both cases, the extent of the matching exercises required to do this is quite considerable. Unmatched data (just the ASC) can be much more easily pooled, enabling analysis of language ethnicity and attainment with larger samples, but without the additional socio-economic variables made possible through the matching exercises. The mapping of languages in *Language Capital* , showing the size and distribution of linguistic groups, provides a ready basis for understanding the potential for pooling.

Of course, the problem with sample size in the example given is that we were, in this instance, primarily interested whether differences in the socio-economic and linguistic characteristics within a broad ethnic group had any relationship with educational outcomes. In other words, did the additional matched data provide any explanations as to the wide spread of outcomes within the Black African group? As part of the project, we also conducted other risk ladder analyses for SEN for all secondary (over 15000) and all primary (over 22000) pupils within Newham. In this case, we included 'Black African' as a 'risk factor' along with the others in the model above. This analysis did produce statistically significant results, and indeed the numbers for secondary and primary cohorts were sufficient to enable comparison between these. For example, we found that eligibility for FSM increased the 'risk' of SEN more at primary level than at secondary level, while conversely being in a single parent household increased the' risk' of SEN at secondary level but not at primary

level. Thus there may certainly well be questions for which risk ladder analysis based on matched ASC and administrative data will generate robust results at the Borough level: we merely caution here on sample size issues in relation to analysis of single year educational attainment and individual language categories, for which pooled data may be necessary.

7. Where to get more information

Enquiries:

Should initially be directed to Dick Wiggins (r.wiggins@ioe.ac.uk) or Michelle vonAhn (Michelle.vonAhn@newham.gov.uk)

Publications

VonAhn, M., Lupton, R., Greenwood, C. and Wiggins R.D. (2010). Languages, ethnicity and Education in London. ESRC UPTAP Research Findings Paper Series., No.10 , July 2010 (http://www.uptap.net/project42.html) . Also available as a Department of Quantitative Social Science Working Paper, No.10-12, June 2010, The Institute of Education, University of London. http://repec.ioe.ac.uk/REPEc/pdf/qsswp1012.pdf

Eversley, J., Mehmedbegović, D., Sanderson, A., Tinsley, T., vonAhn, M. and Wiggins, R.D. (2010). Language Capital: Mapping the languages of London's schoolchildren. Central Books Ltd., London, UK. ISBN-13: 978-904243-95-3 (http://www.cilt.org.uk/shop.aspx).

Presentations

Please see UPTAP website: http://www.uptap.net/project42.html

Websites

London Education Research Unit (LERU) :

> http://www.leru.org.uk/current research collaboration/uptap (languages)/index.html
>
> http://www.leru.org.uk/publications and resources/london maps/index.html

Neighbourhood Knowledge Management/ Mayhew Harper Associates:

> http://www.nkm.org.uk
>
> Also, an interesting case study for risk ladder analysis: http://www.slough.gov.uk/documents/2-LesMayhew.pdf

8. About the Contributors

Michelle von Ahn is Senior Demographic Adviser at the London Borough of Newham.

Ruth Lupton is a Senior Research Fellow at the Centre for Analysis of Social Exclusion at the London School of Economics. She was formerly head of the London Education Research Unit at the Institute of Education.

Richard Wiggins joined the Institute of Education, University of London as Head and Chair of Quantitative Social Science in 2007. His research includes ageing, health and well-being as well as language capital.

John Eversley is Senior Lecturer in City University Community and Health Sciences and Senior Lecturer at London Metropolitan University and managing director of ppre Limited, a not-for- profit social research company.

Antony Sanderson is lead manager for Surrey's education service for equality and language minority achievement. He acted as consultant to the Department for Education in preparing for pupil language data collection through the School Census in England and advised on the launch of the same process in Wales.

Les Mayhew is Managing Director of Mayhew Harper Associates Ltd. and Professor of Statistics at Cass Business School, Faculty of Actuarial Sciences and Insurance.

9. Acknowledgements

Dina Mehmedbegović and Charley Greenwood (LERU, IoE) for their support with and dissemination and data analysis and to CiLT (particularly Teresa Tinsley and Rick Sutton) for their sponsorship and distribution of *Language Capital*. We would also like to thank other members of our advisory group for their guidance on the project: Marnie Caton, (LB Islington), Andrew Cooke (Think London), David Ewens (GLA), Lid King (DcFS), David Lawrence (LSHTM), Jean Martin (ONS, retired), Nicola Morton (London Councils), Kim Price (LB Ealing), Rick Sutton (CiLT), Teresa Tinsley (CiLT). All work funded by the Economic and Social Research Council's Understanding Population Trends and Processes (UPTAP) Grant no. RES-163-27-2004.

References

Anderson. Brown I, Clayton R, Dowty T, Korff, D, Munro, E (2006) *Children's Databases – Safety and Privacy -A Report for the Information Commissioner* Foundation for Information Policy Research. August. http://www.fipr.org/childrens_databases.pdf (Accessed 12 Mar. 10)

Aspinall, P.J (2007) Language ability: A neglected dimension in the profiling of populations and health service users. *Health Education Journal*, Vol. 66, No. 1, 90-106

Baker, P. and Eversley, J. (2000) *Multilingual Capital: the languages of London's schoolchildren and their relevance to economic, social and educational policies.* London: Corporation of London, Battlebridge publications.

CBI (2006) *London Business Survey*

Chriost, Diarmait Mac Giolla and Thomas, Huw(2008) 'Linguistic Diversity and the City: Some Reflections, and a Research Agenda', *International Planning Studies*, 13: 1, 1 — 11

CiLT (2006) *Language Trends 2005: Community Language Learning in England, Wales and Scotlan*d. London: CILT

Crawford, C. and Vignoles, A. (2010) An analysis of the educational progress of children with Special Educational Needs. DOQSS Working Paper 10-19. London: Institute of Education

Dalby, D (1999) *The Linguasphere register of the world's languages and speech communities,* Hebron, Wales: Linguasphere Press

Demie, F and Strand, S (2006). English language acquisition and educational attainment at the end of secondary school .*Educational Studies, Vol. 32, No. 2, , pp. 215–231*

Duckworth, K., Akerman, R., Gutman, L. and Vorhaus, J. (2009) *Influences and leverages on low levels of attainment: a review of literature and policy initiatives.* London: Institute of Education, Centre for Research on the Wider Benefits of Learning

Eversley, J., Mehmedbegovic, D., Sanderson, A., Tinsley, T., VonAhn, M., and Wiggins, R.D. (2010) *Language Capital: Mapping the Languages of London's School Children*. London: CILT

(ICO) Information Commissioner's Office (2010a) http://www.ico.gov.uk/for_organisations/data_protection_guide/list_of_the_data_protection_principles.aspx

(ICO) Information Commissioner's Office (2010b) *Data Protection Technical Guidance Determining what is personal data* http://www.ico.gov.uk/upload/documents/library/data_protection/detailed_specialist_guides/personal_data_flowchart_v1_with_preface001.pdf (Accessed 5 March 2010)

(OPSI) Office of Public Sector Information (2010) *Data Protection Act 1998 Section 33* http://www.opsi.gov.uk/Acts/Acts1998/ukpga_19980029_en_5#pt4-l1g33 (Accessed 4 January 2010)

Ofsted (2008) *Every Language Matters*. London: Ofsted.

Qualifications and Curriculum Authority (QCA) (2006) *Community Languages in the National Curriculum: Report and Guidance*

UKCES (2010) Skills for Jobs Today and Tomorrow: National Strategic Skills Audit for England 2010: Volume 2 The Evidence Report

vonAhn, M., Lupton, R., Greenwood, C., and Wiggins, R. (2010) *Languages, Ethnicity and Education in London.* DoQSS Working Paper No. 10-12 June 2010

APPENDIX 1: Language Categories, School Census

Main Code	Sub-Code	Descriptor
ACL		**Acholi**
ADA		**Adangme**
AFA		**Afar-Saho**
AFK		**Afrikaans**
AKA		**Akan/Twi-Fante**
AKA	**AKAF**	Akan (Fante)
AKA	**AKAT**	Akan (Twi/Asante)
ALB		**Albanian/Shqip**
ALU		**Alur**
AMR		**Amharic**
ARA		**Arabic**
ARA	**ARAA**	Arabic (Any Other)
ARA	**ARAG**	Arabic (Algeria)
ARA	**ARAI**	Arabic (Iraq)
ARA	**ARAM**	Arabic (Morocco)
ARA	**ARAS**	Arabic (Sudan)
ARA	**ARAY**	Arabic (Yemen)
ARM		**Armenian**
ASM		**Assamese**
ASR		**Assyrian/Aramaic**
AYB		**Anyi-Baule**
AYM		**Aymara**
AZE		**Azeri**
BAI		**Bamileke (Any)**
BAL		**Balochi**
BEJ		**Beja/Bedawi**
BEL		**Belarusian**
BEM		**Bemba**
BHO		**Bhojpuri**
BIK		**Bikol**
BLT		**Balti Tibetan**
BMA		**Burmese/Myanma**
BNG		**Bengali**
BNG	**BNGA**	Bengali (Any Other)
BNG	**BNGC**	Bengali (Chittagong/Noakhali)
BNG	**BNGS**	Bengali (Sylheti)
BSL		**British Sign Language**
BSQ		**Basque/Euskara**
BUL		**Bulgarian**
CAM		**Cambodian/Khmer**
CAT		**Catalan**
CCE		**Caribbean Creole English**
CCF		**Caribbean Creole French**
CGA		**Chaga**
CGR		**Chattisgarhi/Khatahi**
CHE		**Chechen**
CHI		**Chinese**
CHI	**CHIA**	Chinese (Any Other)
CHI	**CHIC**	Chinese (Cantonese)
CHI	**CHIH**	Chinese (Hokkien/Fujianese)
CHI	**CHIK**	Chinese (Hakka)
CHI	**CHIM**	Chinese (Mandarin/Putonghua)
CKW		**Chokwe**
CRN		**Cornish**
CTR		**Chitrali/Khowar**
CWA		**Chichewa/Nyanja**

Main Code	Sub-Code	Descriptor
CYM		**Welsh/Cymraeg**
CZE		**Czech**
DAN		**Danish**
DGA		**Dagaare**
DGB		**Dagbane**
DIN		**Dinka/Jieng**
DUT		**Dutch/Flemish**
DZO		**Dzongkha/Bhutanese**
EBI		**Ebira**
EDO		**Edo/Bini**
EFI		**Efik-Ibibio**
ENB		**Believed to be English***
ENG		**English***
ESA		**Esan/Ishan**
EST		**Estonian**
EWE		**Ewe**
EWO		**Ewondo**
FAN		**Fang**
FIJ		**Fijian**
FIN		**Finnish**
FON		**Fon**
FRN		**French**
FUL		**Fula/Fulfulde-Pulaar**
GAA		**Ga**
GAE		**Gaelic/Irish**
GAL		**Gaelic (Scotland)**
GEO		**Georgian**
GER		**German**
GGO		**Gogo/Chigogo**

Main Code	Sub-Code	Descriptor
GKY		**Kikuyu/Gikuyu**
GLG		**Galician/Galego**
GRE		**Greek**
GRE	**GREA**	Greek (Any Other)
GRE	**GREC**	Greek (Cyprus)
GRN		**Guarani**
GUJ		**Gujarati**
GUN		**Gurenne/Frafra**
GUR		**Gurma**
HAU		**Hausa**
HDK		**Hindko**
HEB		**Hebrew**
HER		**Herero**
HGR		**Hungarian**
HIN		**Hindi**
IBA		**Iban**
IDM		**Idoma**
IGA		**Igala**
IGB		**Igbo**
IJO		**Ijo (Any)**
ILO		**Ilokano**
ISK		**Itsekiri**
ISL		**Icelandic**
ITA		**Italian**
ITA	**ITAA**	Italian (Any Other)
ITA	**ITAN**	Italian (Napoletan)
ITA	**ITAS**	Italian (Sicilian)
JAV		**Javanese**
JIN		**Jinghpaw/Kachin**

Main Code	Sub-Code	Descriptor
JPN		**Japanese**
KAM		**Kikamba**
KAN		**Kannada**
KAR		**Karen (Any)**
KAS		**Kashmiri**
KAU		**Kanuri**
KAZ		**Kazakh**
KCH		**Katchi**
KGZ		**Kirghiz/Kyrgyz**
KHA		**Khasi**
KHY		**Kihaya/Luziba**
KIN		**Kinyarwanda**
KIR		**Kirundi**
KIS		**Kisi (West Africa)**
KLN		**Kalenjin**
KMB		**Kimbundu**
KME		**Kimeru**
KNK		**Konkani**
KNY		**Kinyakyusa-Ngonde**
KON		**Kikongo**
KOR		**Korean**
KPE		**Kpelle**
KRI		**Krio**
KRU		**Kru (Any)**
KSI		**Kisii/Ekegusii (Kenya)**
KSU		**Kisukuma**
KUR		**Kurdish**
KUR	**KURA**	Kurdish (Any Other)
KUR	**KURM**	Kurdish (Kurmanji)
KUR	**KURS**	Kurdish (Sorani)
LAO		**Lao**
LBA		**Luba**
LBA	**LBAC**	Luba (Chiluba/Tshiluba)
LBA	**LBAK**	Luba (Kiluba)
LGA		**Luganda**
LGB		**Lugbara**
LGS		**Lugisu/Lumasaba**
LIN		**Lingala**
LIT		**Lithuanian**
LNG		**Lango (Uganda)**
LOZ		**Lozi/Silozi**
LSO		**Lusoga**
LTV		**Latvian**
LTZ		**Luxemburgish**
LUE		**Luvale/Luena**
LUN		**Lunda**
LUO		**Luo (Kenya/Tanzania)**
LUY		**Luhya (Any)**
MAG		**Magahi**
MAI		**Maithili**
MAK		**Makua**
MAN		**Manding/Malinke**
MAN	**MANA**	Manding/Malinke (Any Other)
MAN	**MANB**	Bambara
MAN	**MANJ**	Dyula/Jula
MAO		**Maori**
MAR		**Marathi**
MAS		**Maasai**

Main Code	Sub-Code	Descriptor
MDV		**Maldivian/Dhivehi**
MEN		**Mende**
MKD		**Macedonian**
MLG		**Malagasy**
MLM		**Malayalam**
MLT		**Maltese**
MLY		**Malay/Indonesian**
MLY	**MLYA**	Malay (Any Other)
MLY	**MLYI**	Indonesian/Bahasa Indonesia
MNA		**Magindanao-Maranao**
MNG		**Mongolian (Khalkha)**
MNX		**Manx Gaelic**
MOR		**Moore/Mossi**
MSC		**Mauritian/Seychelles Creole**
MUN		**Munda (Any)**
MYA		**Maya (Any)**
NAH		**Nahuatl/Mexicano**
NAM		**Nama/Damara**
NBN		**Nubian (Any)**
NDB		**Ndebele**
NDB	**NDBS**	Ndebele (South Africa)
NDB	**NDBZ**	Ndebele (Zimbabwe)
NEP		**Nepali**
NOR		**Norwegian**
NOT		**Information not obtained***
NUE		**Nuer/Naadh**
NUP		**Nupe**
NWA		**Newari**
NZM		**Nzema**
OAM		**Ambo/Oshiwambo**
OAM	**OAMK**	Ambo (Kwanyama)
OAM	**OAMN**	Ambo (Ndonga)
OGN		**Ogoni (Any)**
ORI		**Oriya**
ORM		**Oromo**
OTB		**Believed to be Other than English***
OTH		**Other than English***
OTL		**Other Language**
PAG		**Pangasinan**
PAM		**Pampangan**
PAT		**Pashto/Pakhto**
PHA		**Pahari/Himachali (India)**
PHR		**Pahari (Pakistan)**
PNJ		**Panjabi**
PNJ	**PNJA**	Panjabi (Any Other)
PNJ	**PNJG**	Panjabi (Gurmukhi)
PNJ	**PNJM**	Panjabi (Mirpuri)
PNJ	**PNJP**	Panjabi (Pothwari)
POL		**Polish**
POR		**Portuguese**
POR	**PORA**	Portuguese (Any Other)
POR	**PORB**	Portuguese (Brazil)
PRS		**Persian/Farsi**
PRS	**PRSA**	Farsi/Persian (Any Other)
PRS	**PRSD**	Dari Persian
PRS	**PRST**	Tajiki Persian
QUE		**Quechua**
RAJ		**Rajasthani/Marwari**

Main Code	Sub-Code	Descriptor
REF		**Refused***
RME		**Romany/English Romanes**
RMI		**Romani (International)**
RMN		**Romanian**
RMN	**RMNM**	Romanian (Moldova)
RMN	**RMNR**	Romanian (Romania)
RMS		**Romansch**
RNY		**Runyakitara**
RNY	**RNYN**	Runyankore-Ruchiga
RNY	**RNYO**	Runyoro-Rutooro
RUS		**Russian**
SAM		**Samoan**
SCB		**Serbian/Croatian/Bosnian**
SCB	**SCBB**	Bosnian
SCB	**SCBC**	Croatian
SCB	**SCBS**	Serbian
SCO		**Scots**
SHL		**Shilluk/Cholo**
SHO		**Shona**
SID		**Sidamo**
SIO		**Sign Language (Other)**
SLO		**Slovak**
SLV		**Slovenian**
SND		**Sindhi**
SNG		**Sango**
SNH		**Sinhala**
SOM		**Somali**
SPA		**Spanish**
SRD		**Sardinian**

Main Code	Sub-Code	Descriptor
SRK		**Siraiki**
SSO		**Sotho/Sesotho**
SSO	**SSOO**	Sotho/Sesotho (Southern)
SSO	**SSOT**	Sotho/Sesotho (Northern)
SSW		**Swazi/Siswati**
STS		**Tswana/Setswana**
SUN		**Sundanese**
SWA		**Swahili/Kiswahili**
SWA	**SWAA**	Swahili (Any Other)
SWA	**SWAC**	Comorian Swahili
SWA	**SWAK**	Swahili (Kingwana)
SWA	**SWAM**	Swahili (Brava/Mwiini)
SWA	**SWAT**	Swahili (Bajuni/Tikuu)
SWE		**Swedish**
TAM		**Tamil**
TEL		**Telugu**
TEM		**Temne**
TES		**Teso/Ateso**
TGE		**Tigre**
TGL		**Tagalog/Filipino**
TGL	**TGLF**	Filipino
TGL	**TGLG**	Tagalog
TGR		**Tigrinya**
THA		**Thai**
TIB		**Tibetan**
TIV		**Tiv**
TMZ		**Berber/Tamazight**
TMZ	**TMZA**	Berber/Tamazight (Any Other)
TMZ	**TMZK**	Berber/Tamazight (Kabyle)

Main Code	Sub-Code	Descriptor
TMZ	**TMZT**	Berber (Tamashek)
TNG		**Tonga/Chitonga (Zambia)**
TON		**Tongan (Oceania)**
TPI		**Tok Pisin**
TRI		**Traveller Irish/Shelta**
TSO		**Tsonga**
TUK		**Turkmen**
TUL		**Tulu**
TUM		**Tumbuka**
TUR		**Turkish**
UKR		**Ukrainian**
UMB		**Umbundu**
URD		**Urdu**
URH		**Urhobo-Isoko**
UYG		**Uyghur**
UZB		**Uzbek**
VEN		**Venda**
VIE		**Vietnamese**
VSY		**Visayan/Bisaya**
VSY	**VSYA**	Visayan/Bisaya (Any Other)
VSY	**VSYH**	Hiligaynon
VSY	**VSYS**	Cebuano/Sugbuanon
VSY	**VSYW**	Waray/Binisaya
WAP		**Wa-Paraok (South-East Asia)**
WCP		**West-African Creole Portuguese**
WOL		**Wolof**
WPE		**West-African Pidgin English**
XHO		**Xhosa**
YAO		**Yao/Chiyao (East Africa)**
YDI		**Yiddish**
YOR		**Yoruba**
ZND		**Zande**
ZUL		**Zulu**
ZZZ		**Classification Pending**