Name recognition systems and self-identification: which predicts education attainment better?

1. Context

There is no shortage of evidence to show that levels of educational attainment vary considerably as between pupils of white British and other ethnic heritages.

Assigning an ethnic group to a pupil is therefore essential for understanding what these differences are. However, as we will demonstrate, the question of how best do this may not be as self-evident as it sometimes seems.

As things stand the current 'best practice' within government involves requiring service-users or survey respondents to specify from a given list the ethnic group to which they belong. Government guidance to schools on the collection of ethnicity data states that:[1]

> "The school must not ascribe any ethnicity to the pupil. This information must come from the parent/ guardian or pupil. Where the ethnicity has not yet been collected, this is recorded as 'NOBT' (information not yet obtained)."

Such an approach is appropriate for informing the treatment of individual pupils. But that does not necessarily mean it is the most effective method of classification where the objective is to quantify disparities in outcomes between ethnic groups.

2. About Origins

Origins is a name recognition tool, developed by Webber Phillips. It is used across the public, private and third sectors, including by a number of government agencies. It is built upon a database of 1.3m unique forenames and 4.0m unique surnames. In essence, Origins allows you to infer the ethnic makeup of a population directly, using the combination of personal and family names.

This paper compares the performance of current 'best practice' with that of Origins, as predictors of pupils' key stage 2 test performance in English and Mathematics in a diverse London borough.

(Appendix A has further details about the methodology. Appendix B looks at some of the political and methodological questions behind the approach).

3. Methodology

This study is based on analysis of the academic performance of 2,148 pupils. The results are based on results for 2018.

The best practice approach uses the 28 ethnicity codes by which the borough's pupils have been coded based on pupil or parental responses. These are referred to as 'Approved extended categories' and can be combined to match the 20 'DfE main codes'.[2]

In addition, the borough used the Origins software to append an Origins code to each pupil based on their first and last names. Once the Origins codes had been added, the names (and any other characteristics that might lead to the identification of a pupil) were removed.

[1] https://www.gov.uk/guidance/complete-the-school-census/data-items
[2] https://www.gov.uk/guidance/complete-the-school-census/find-a-school-census-code

For the purpose of this analysis the ethnicity of the pupils in the database are coded in two ways: firstly into the 20 DfE categories used by schools and the education system more generally, and secondly into the 50 Origins sub-groups. We have then used performance in English and Mathematics to understand which system for classifying by ethnicity reveals greater variation in attainment (i.e. higher 'predictiveness').

4. Measuring predictiveness

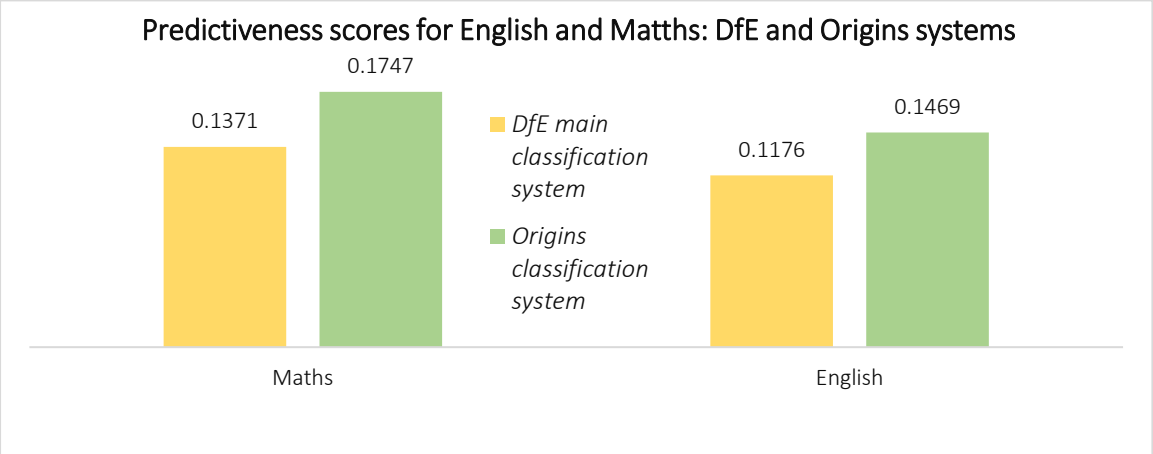
By 'predictiveness' we refer to the strength of the relationship between ethnicity and attainment. The stronger the relationship, the better the one is able to predict the other. We quantified the predictiveness of the two ethnic coding systems using the following process.

- *Firstly,* we calculated, for each ethnic category, the average Key Stage 2 score among those who took tests;
- *Secondly,* we measured the difference between the average score for pupils in each ethnic category and the average score for all pupils (i.e. the deviation from the overall average of pupils in each ethnic group);
- *Thirdly,* we calculated the average difference across all ethnic categories, weighting on the basis of the number of pupils in each ethnic group; this allowed us to create a 'predictiveness score', which essentially acted us a top-line measure of the strength of the link between ethnicity and educational outcome based, on each of the two different ethnic coding systems;
- *Lastly,* we compared the predictiveness scores for the DfE codes with those for Origins, in English and Mathematics respectively.

(Appendix C looks into the specific methodological question of the number of codes used by the respective systems).

5. Results

The exercise revealed some interesting variations, such as the strong performance of pupils of Albanian heritage, especially compared with that of pupils from neighbouring Former Yugoslav nations. Likewise, it revealed the high attainment of children from south Asian backgrounds, especially in Mathematics. These differences were identified by both coding systems.



More striking, however, was the overall difference between the predictiveness score of the two classification systems. The chart above shows that, in the case of both English and Mathematics, Origins' predictiveness scores were significantly higher than the DfE's classification. (Maximum

possible predictiveness would be 1.0 in the case of each). Predictiveness scores were 13.8% greater in the case of English and 21.6% greater in the case of Mathematics.

Why should a pupil's name be more predictive of their attainment than the ethnicity assigned to them by their parents? One possible reason is that there are more Origins codes than there are DfE codes. However when we repeated the analysis but just for categories with nine or more pupils, we found the advantage of Origins over the DfR classification even stronger.

Perhaps the most obvious explanation is how Origins breaks down the very large 'white other' category into its behaviourally very different components. Another is that, whilst the DfE classification can often be interpreted in terms of nationality, names provide a surer indicator of heritage. For example pupils with Albanian names may originate from Kosovo as well as from Albania, or from families previously resident in Germany or Switzerland.

Other factors relate to trust and identity. Our analysis in April of data for hospital admissions within an NHS Trust showed a strong proclivity of respondents with non-white and non-British names to identify as 'British' or as 'white' in terms of ethnicity. The same study revealed a disproportionately high number of those with non-white and non-British names being coded as 'not known' or not being coded at all.[3]

6. Conclusions

One cannot say which of the two systems we have tested is the more 'accurate' as a means of measuring of an individual's ethnicity. Identity is, in many respects, subjective and depends on circumstances. However, it is possible to say which system is more predictive of outcomes and, if inequalities of outcome are a source of concern to policy makers, then it is logical to focus on whichever classification better predicts inequality.

As our chart above shows, the Origins method of data collection does reveal larger variations in attainment and a stronger correlation between ethnicity and outcome than self-completion does. The tool is clearly spotting something which conventional methods are not. Hence, there is a real risk that the reliance on self-identification for statistical analysis is dampening the extent to which ethnic disparities are recognised.

[3] https://webberphillips.com/wp-content/uploads/2021/05/How-accurate-is-NHS-ethnicity-data-REPORT-April-2021.pdf

Appendix A: more detail on Origins

The attribution of names varies between cultures – sometimes occurring based on national borders, sometimes by faith and sometimes by language. Origins identifies 11 core categories (e.g. East European names), 50 sub-categories of names (e.g. Baltic names) and about 200 specific classifications (e.g. Lithuanian names).

The tool analyses both forenames and surnames, attributing a confidence score for each and weighting certain categories of names. It has been tested extensively and is able to assess, with a high degree of granularity, the ethno-cultural make-up of a given group of names.[4] The larger the sample, the higher the degree of accuracy.

The use of names, meanwhile, allows decision-makers to obtain information for more discrete categories. We can identify the proportion of residents that are ethnically Albanian, for instance – rather than relying on the nebulous census category 'white other'.

The advantage of the tool is thus that it is able to offer greater coverage, consistency and ease than regular sampling. Analysis can be carried out at LSOA or postcode level, where needed.

Clearly no predictive system is likely to be able to predict the ethnicity of an adult with 100% accuracy from their name. However at the level of detail such as 'South Asian', 'Black African', 'Eastern European' or 'Turkish' the accuracy is very high.

We access commercially available data and weight it based on ONS population estimates, to look at the makeup of Britain's adult population. This allows us a comparison of the personal and family names of virtually all adults living in Britain in March 2011 and March 2021.

Clearly the heritage of not every British adult can be accurately inferred from their name, but overall if the percentage of British adults with Romanian names has increased three-fold over this period then it is reasonable to infer that the proportion of British adults of Romanian heritage is likely to have increased by a not dissimilar amount.

Appendix B: Political and methodological considerations

Opinion regarding the use of names to infer ethnic origin often focuses on the extent to which name-based ethnicity attribution systems are 'accurate'. The use of the term 'accuracy' in turn supposes first that there is an unambiguous ethnic classification that applies to any individual and second that this classification can be elicited through self-identification.

For example, the footballer Mesut Ozil plays international football for Germany, where he was born, but his name betrays his Turkish ancestry. In terms of his behaviour it may well be that he identifies as a German on the football pitch but as Turkish when he chooses which restaurant to visit. There are likely to be elements of both ethnic groups in his behaviour.

[4] In order to comply with GDPR regulations and current COPI protocols Origins codes are only held against anonymised records. In other words no inferred ethnicity data is held against recognisable individuals.

Rather than debate 'how accurate' the Origins software is we consider it more useful to ask ourselves the question 'is a person's name a better predictor of their behaviour than their self-declared ethnicity?'.

Appendix C: The number of categories

It might be thought that the higher predictiveness score achieved by Origins may have arisen from its having a greater number of categories than the DfE classification. On this basis it could be argued that the comparison is unfair.

In light of these considerations, we have done further analysis to compare the predictiveness of the two classifications, by restricting the comparison of their effectiveness to just those categories with more than 8 pupils. This helps to show whether Origins' improved predictiveness was the result of more accurate categorisation or of it revealing a larger number of categories.

By applying this restriction we find that the Origins classification retains 77% of its predictive power and the DfE one 61% of its predictive power. In other words it is not the greater number of Origins categories which contributes to its superior predictiveness.