Young participation in higher education: England to learn from Wales, Northern Ireland and Scotland

Comments on January 2005 HEFCE report

Understanding variation in participation rates is important. That is why I am appalled by the January 2005 HEFCE study: any statistician would have known the study was doomed to be useless before it started.

When commissioning research, you must begin by defining the question to be answered. This question, and its sub-questions must be clearly stated. There is no point in wasting large sums of money on collecting or collating data without considering how the results might be used. Therefore, with respect to participation in higher education, merely asking how area of residence and participation are related over seven years, during which many changes were made to school examinations, is silly. The study design does consider birth month and sex, univariately, but ignores the major predictor of participation: schooling. A token acknowledgment is made by considering the 'local' school. There are numerous remarks in the report about having sufficiently accurate measures, but there is no evidence that any thought was given to the appropriate sample size.

Two examples illustrate the problem. In public health, a study which proposed to report death statistics on heart attacks by social class while ignoring the major confounders of smoking, alcohol use and body weight would be recognized as ridiculous. In agriculture, trying to evaluate fertilizers by measuring the yields of wheat from fields in Sutherland, in Cornwall and Essex while ignoring the effects of the rainfall, sunshine, soil and latitude was abandoned over 150 years ago.

It seems that ignoring major sources of variation is regard as good practice by HEFCE. Why bother learning that premature heart attacks can be reduced by sensible early actions if you can blame the geographical distribution of general practices?

I regard research which cannot deliver useful information as unethical, as it wastes money and time. It is even worse if the conclusions drawn are misleading. I searched the HEFCE website, but failed to find anything on research ethics. Under 'Good practice', universities are given instructions on procurement, but HEFCE does not appear to have a policy on commissioning research. Under 'Fraud' there is reference to leadership and commercial ethics. I don't find evidence of leadership or ethical use of funds in this study and report. At no point does the report state a clear aim or aims. Even in an exploratory study, it is inexcusable to ignore the major confounders. There are some good points in the report. The issues of double counting, missing information, definition of area of residence at age 16 and control groups are explored.

Conclusions drawn

Assume for the moment that the numerical results are correct (this assumption is discussed later).

The missing major conclusion is that England has a lot to learn from Wales, Northern

Ireland and Scotland, which all have higher participation rates, Scotland much higher. Scotland has less inequality in participation ratios than England: Figure 31(d) shows YPR(A) ratio for England ranged from 4.9 to 4.6, Figure 32(d) that the Scottish ratios were 3.5 to 3.9. Why was this not stated?

Although all the results in section 3 show a reduction in inequality, with the ratios decreasing, the summary claims 'Participation inequality between neighbourhoods persistent'. Despite the repeated claims of accuracy, the report gives no information on the accuracy of the ratios (the remark about 'errors in the estimates, of 1-3 percent' in 3.26 is too vague to be of use). However, the authors comment on the persistence of the general trends. The reader can consider whether if there is no underlying change, is it likely that 12 out of 12 results are positive.

No comment is made on what a plausible change in participation ratios might be, but the changes will be restricted by the size of the school cohort, number of places available, and the aggregation chosen. A decrease in partition ratio using census wards from 6.0 to 5.5 over seven years (Figure 29(d)) might be as good as it gets. The different ratios are partly a function of the size of area underlying the formation of the quintiles.

'Deep divisions in the participation chances of young people by where they live' is misleading, as it implies that area of residence, rather than several factors, such as sex, month of birth, schooling and socio-economic status, is responsible for the divisions.

The report fails to use the data effectively: in Scotland, children start school 6 months older. The pattern of seasonal variation in participation is different from England: the oldest children do not have the highest participation rate. Instead of using the differences in school year in Scotland to consider the difference between being youngest in the year, and starting school at either age 4 or age 4.5, there is a perjorative reference to the 'problematic school year structure in Scotland'. As Scotland has much higher participation rates, should the reference not be to the problematic school year structure in England?

I would like to draw attention to one way in which the English education policy systematically discriminates against children from lower socio-economic groups. There is substantial variation in entry by month of birth, with the youngest children being less likely to go on to higher or further education. Now, premature birth has a social class gradient: proportionally more children in dis-advantaged areas are premature than in advantaged areas. Children who are born prematurely in July or August have to start school a year earlier than if they were born full term. Such children will on average be smaller and clumsier, as well as the youngest in the year, and children from dis-advantaged backgrounds will be over-represented in this group. It would be nice to see some recognition of prematurity in statutory start dates.

Errors in methods used

Doubt as to the correctness and competence of the analysis exists beyond a reasonable level. You notice if someone refers to the sun as 'she'; to write 'the female uterus' would indicate not an error of fact, but an error of comprehension. A statistician notices the phrase 'discrete binomial distribution'. Some of the other errors are mentioned here. The methods used in the analysis are at least 15 years out of date. Binary data of this kind can and should be analysed using generalised linear models and their extensions. Analysis of binary data at individual level has been done for at least 50 years: it is not necessary to use grouped data. Annex F shows a failure to understand the difference between systematic and random variation. With the correct statistical methods, there is no tension between using smaller areas 'to better target participation disadvantage' and wanting large areas

to avoid 'large random fluctuations in the observed participation rates'. Despite having stated the concept of underlying propensity, the authors claim that data for a cohort of size five 'is far too variable for any useful area classification' as fluctuations give rates of zero or over 60%. But conclusions about the effect of covariates can be reliably drawn using binary data, where the only possible rates are zero and 100%.

Annex A.6 demonstrates the lack of basic understanding of statistics: 'using two regression models (least squares ....) and Poisson regression'. 'Least squares' is not a regression model, it is an optimization criterion. Formula 1 fails to define a model. The term $e_{y,i}$ is not defined, and I'd be curious to know what definition was used in the Poisson regression. As a separate model is fitted for each of five years, the 'y' is redundant, except to hint at the correct method for assessing linearity of parameter estimates and year of estimate. The authors fail to report standard errors for the parameter estimates. In Annex B.2, it is stated that liner extrapolation may not be suitable, but it was still used. Figure 69 demonstrates the that the choice of estimate affects the participation ratio by as much as 5%: ratio of 4.94 versus 4.72.

Annex F.2 demonstrates profound ignorance of real and effective sample sizes: anyone analysing hierarchical or clustered data should be able to partition the 'random fluctuations' into components. Heterogeneity is problematic if the causes of heterogeneity are unknown. When the sources of heterogeneity are ignored, the problem is the analysis, not the heterogeneity. Annex F.3 (see Table 13) fails to give a proper description of the simulation experiment, and ignores the statistical literature on extra-Binomial variation.

Good practice in study design

Research, or even 'audit' which fails to take into account previous work in the study design, execution and analysis is bad. Anyone who has a slight knowledge of higher education knows that entry at 18 or 19 is strongly related to achievement in public examinations. Achievement in public examinations is known to vary considerably within and between schools. The government promotes league tables of school achievement. Variation by month of birth and sex is also familiar.

As the study design uses UCAS data, the cost of including individual student level information on school type and school results for those who participate should not be large. The costs of collecting relevant information on those who do not enter the UCAS system is likely to be higher. However, without this information, the effects of schooling cannot be distinguished from the effects of area of resident, age or sex. As schooling until 16 years mandatory, lists of all children aged 15-16, in their last compulsory year, and their GCSE results are available: such data are required to provide the school league tables. Schools and colleges will have data on their scholars, and those who have dropped out could be initially identified by exclusion. At least these data should be linked to young participation. There is no excuse for carrying out a large but uninformative census instead of a properly designed sample. For the same expenditure, it should be possible to obtain far more useful information.

A properly designed study addresses the question of how much data is required to address the primary research aims. Justification of sample size is essential in obtaining ethics permission for medical and health studies.

As the report does, I first ignore the relevant covariates. Of course, a sensible approach to study size would consider the variation in participation after adjusting for the known covariates on schooling, sex, and birth month. The report states that likely annual changes will be small (1.5%), but never states what size of difference is worth measuring, or what

will be deemed 'no change.' In Annex A errors of sizes 0.2%, 1%, 1.5% and 3% are mentioned. Biases of the order of order of 5% and 2% are discussed in Annex B.2. With the English cohorts, a quintile has 100,000 young adults, and hence percentages, say 10% or 50% are measured with standard errors of 0.09% and 0.16%. In other words, the standard errors are smaller than is justified by the data quality.

In the light of the serious shortage of quantitative skills in the social sciences, it is not surprising that poor design and analysis of large studies occurs. The First Report of the Select Committee on Science and Technology [1] expressed deep concern at these shortages.

The Statistics Commission is an independent public body, which aims to help improve the quality and relevance of official statistics by identifying areas that require improvement [2]. If participation statistics are official statistics, then HEFCE could request assistance from the Statistics Commission.

Implications for fair access

I am concerned by the statements by Sir Martin Harris of Offa: 'the HEFCE research was "by far the most comprehensive and detailed analysis of its kind. It demonstrates there is still more work to be done to ensure equality of opportunity for those students from low-income groups" '. This conclusion is untenable. To reach this conclusion from *this* study requires the joint assumption that innate ability to benefit from education is completely independent of socio-economic circumstances and that equality of opportunity is essentially perfect for school education. The first assumption is implied in the report to be false: better education is assumed to lead to better employment and socio-economic circumstances, and no evidence is adduced that ability is not partly inherited from parents. With respect to the second assumption, consider the last sentence of the summary: 'This suggests that where you lived as a child, so important in determining earlier educational outcomes, has little additional effects (sic) on the transition to *postgraduate* study.' But this report *ignores* 'earlier educational outcomes'.

This study could not demonstrate equality of opportunity of access of higher education, as there is no information on individual's earlier education. The results reported are consistent with people from fee-paying schools with good A-levels being less likely to be accepted than those from state schools with the same, or poorer A-levels. (See 'Bristol admissions secrets revealed' A. Goddard, The Times Higher Feb 4 2005) You might believe there is the opposite inequality, but this report provides no data to discriminate between the two views.

It might well be that if school type and results were properly taken into account, the conclusions might be that it is school education which needs to be addressed, and that higher and further education establishments are needlessly blamed for the failings of compulsory education.

A government which had confidence in the provision of school education would have no need for an Office of Fair Access. I welcome the recent Labour and Tory comments on school discipline, as this will possibly do more for equality of educational opportunity than creating sticks with which to beat universities.

Professor J L Hutton,
Department of Statistics, University of Warwick. February 9, 2005

[1] http://www.publications.parliament.uk/pa/cm200405/cmselect/cmsctech/13/1302 .htm
[2] http://www.statscom.org.uk/index.asp