Department of Quantitative Social Science

Non-response biases in surveys of school children: the case of the English PISA samples

John Micklewright
Sylke V. Schnepf
Chris Skinner

DoQSS Working Paper No. 10-04
June 2010

Leading education and social research
Institute of Education
University of London

DISCLAIMER

Any opinions expressed here are those of the author(s) and not those of the Institute of Education. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

DoQSS Workings Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Department of Quantitative Social Science. Institute of Education, University of London. 20 Bedford way, London WC1H 0AL, UK.

Non-response biases in surveys of school children: the case of the English PISA samples

John Micklewright[*], Sylke V. Schnepf[†], Chris Skinner[‡§]

Abstract. We analyse response patterns to an important survey of school children, exploiting rich auxiliary information on respondents and non-respondents cognitive ability that is correlated both with response and the learning achievement that the survey aims to measure. The survey is the Programme for International Student Assessment (PISA), which sets response thresholds in an attempt to control data quality. We analyse the case of England for 2000 when response rates were deemed high enough by the PISA organisers to publish the results, and 2003, when response rates were a little lower and deemed of sufficient concern for the results not to be published. We construct weights that account for the pattern of non-response using two methods, propensity scores and the GREG estimator. There is clear evidence of biases, but there is no indication that the slightly higher response rates in 2000 were associated with higher quality data. This underlines the danger of using response rate thresholds as a guide to data quality.

JEL classification: C83, I21.

Keywords: non-response, bias, school survey, data linkage, PISA.

[*]Department of Quantitative Social Science, Institute of Education, University of London. 20 Bedford Way, London WC1H 0AL, UK. E-mail: (J.Micklewright@ioe.ac.uk)

[†]School of Social Sciences, University of Southampton, UK. E-mail: (svs@soton.ac.uk)

[‡]School of Social Sciences, University of Southampton, UK. E-mail: (C.J.Skinner@soton.ac.uk)

[§]**Acknowledgements**. This research was funded by ESRC grant RES-062-23-0458 ('Hierarchical analysis of unit non-response in sample surveys') and was also supported by grant RES-576-25-0014 (the ADMIN node of NCRM). It builds on earlier work funded by the (then) Department for Education and Skills (DfES). We thank staff at DfES, the Office for National Statistics, and the Fischer Family Trust for their help in setting up and interpreting the data. We alone remain responsible for their use. We are grateful for comments to seminar and conference participants at DfES, the ESRC Research Methods Festival, the European University Institute, Southampton, and Bremen. Micklewright thanks the Institute for Research on Poverty at the University of Wisconsin-Madison for hospitality during a sabbatical visit when he worked on the paper.

1. Introduction

Surveys of school children are carried out in many industrialized countries as a result of national debates about education policy and as a part of international inquiries into student performance. Potential bias from non-response represents a major threat to the validity of findings from such surveys. A common approach adopted by survey organizers or funders to maintain data quality in the face of non-response is to require response rates to exceed a target threshold. For example, thresholds of 85 percent for school response and 80 percent for student response are set in the Programme for International Student Assessment (PISA), co-ordinated by the OECD. The Trends in International Maths and Science Study (TIMSS) seeks response rates of 85 percent for both schools and students.

Such thresholds provide an appealing rule of thumb but they are no guarantee: the *pattern* of response needs to be considered and not just the rate (e.g. Groves 1989, 2006). Low response may result in little bias if respondents and non-respondents are similar. High response may be consistent with non-trivial bias if the characteristics of those not responding are very different. In contrast to the relative simplicity with which response rates may be measured, assessing non-response bias usually represents a difficult challenge, since information about non-respondents is often very limited. What is needed is comparable information on the characteristics of respondents and non-respondents that is both correlated with response *and* correlated with key survey outcome variables.

This paper exploits rich auxiliary information on respondents and non-respondents to one survey that does satisfy these conditions. Our aim is to analyse non-response biases in England to the first two rounds of PISA, which began in 2000 and is conducted every three years. We have individual level microdata on cognitive achievement for the entire population of 15 year children in schools from which the PISA sample is drawn. We are able to link this information to the PISA sample. This is a very unusual situation – we have information for both respondents and non-respondents and the rest of the population in exactly the subject area that is the focus of the survey.

It is especially important to consider the English sample in PISA. Reports from OECD for the 2003 round excluded the UK following concerns that the quality of data for England, where there is a separate survey, suffers from non-response bias.

Not surprisingly, this was the subject of considerable comment. For example, speaking at the 2005 Royal Statistical Society Cathie Marsh lecture on 'Public Confidence in Official Statistics', Simon Briscoe of *The Financial Times* listed this incident as among the 'Top 20' recent threats to public trust in official statistics. We also estimate the extent of biases in the 2000 data when response rates at both school and pupil levels in England were little higher than in 2003. As for other countries, the microdata for England for both rounds can be downloaded from the PISA website (www.pisa.oecd.org). The data are therefore available for use worldwide, underlying the importance of research into their quality.

There has long been a need to obtain a better understanding of response to school surveys in England. Relative to other countries, England has had a poor record in the international surveys of children's cognitive achievement, including TIMSS and the Progress in Reading Literacy Study (PILRS) as well as PISA. For example, the average response rate for OECD countries in eight surveys between 1995 and 2003 – TIMSS 1995 4th grade, TIMSS 1995 8th grade, TIMSS 1999, PISA 2000, PIRLS 2001, TIMSS 2003 4th grade, TIMSS 2003 8th grade, and PISA 2003 – exceeded those for England by 30 percentage points for school response 'before replacement', 12 points for school response 'after replacement' (these terms are defined below) and 5 points for pupil response. Moreover, response rates to school surveys organised by government have fallen over time, by an estimated average of about 2 percentage points per year over 1995-2004 (Sturgis et al. 2006, analysis of 73 surveys). Happily, in the case of PISA, response in England was substantially higher in 2006 and results for the UK were included back into the OECD's reports. But the uncertainty about data quality in 2000 and 2003 remains and higher response in 2006 does not imply that any problems were absent.

Section 2 summarises the PISA sample design and response in England in 2000 and 2003. It also describes our auxiliary information, which is drawn from administrative registers on pupil performance in national tests taken at age 14 and in public exams taken shortly after the PISA fieldwork period.

Response patterns in a survey may result in biases in estimates of some parameters of interest but not others. Section 3 analyses the test and exam scores to assess biases in estimates of: (a) mean achievement, (b) dispersion of achievement, and (c) the percentage of children below a given achievement threshold. These measures summarise the main features of interest of the distribution: how well children are doing on average, the differences among them, and the numbers not meeting particular standards. We show that biases arise mainly from pupil response rather than school response, especially in 2003, and then provide further analysis of the pupil response probability using logistic regression models.

Section 4 uses two methods to construct alternative sets of weights to adjust for the pattern of response. The first uses propensity scores based on results from the logistic regression models in Section 3. The second method exploits the fact that we have auxiliary information for the entire target population. We estimate weights based on the generalised regression (GREG) estimator, weights that account for differences between the composition of the achieved sample of responding pupils and the composition of the population from which they were drawn.

In Section 5 we apply these alternative sets of weights to the sample of respondents. The focus is now on estimates of achievement based on PISA test scores. We again consider the three parameters of the distribution described above. We also briefly consider regression parameters describing the association between achievement and measures of family background that are recorded in the survey data. In each case, a comparison of the results with those obtained when we use the survey design weights provides estimates of the extent of non-response bias. Section 6 discusses the results within the paradigm of total survey error. Section 7 concludes.

2. PISA data for England and the auxiliary information

Sample design

PISA has a two stage design. First, schools with 15 year olds are sampled with probability proportional to size (PPS). Second, a simple random sample of 35 pupils aged 15 is drawn for each responding school. If a school has fewer than 35 pupils of this age, all are included in the second stage sample. Pupil sampling is done by the survey agency, the Office for National Statistics (ONS) in 2000 and 2003, from lists provided by schools.

The first stage sampling is stratified on school size and type of school – state, quasi-state, and private (the English terms for these three types are 'LEA', 'grant maintained' and 'independent'). The great majority of schools are state schools and only 7 percent of 15 year olds attend private schools. Within the first two types, further strata are created on the basis of region and, importantly, the age 16 public exam records that form part of our auxiliary information. Private schools are stratified by region and by pupil gender.

As is common with the international surveys on children's learning, a system of 'replacement' of non-responding schools is operated. Replacement in survey sampling is the subject of considerable debate (e.g. Vehovar 1999, Prais 2003, Adams 2003, Lynn 2004, Sturgis et al. 2006). The PPS sampling generates a list of 'initial schools' together with two potential replacements, the schools that come immediately before and after within the stratum. If an initial school declines to participate, its first replacement is approached. If this first replacement does not respond, the second replacement is asked to participate. Schools sampled in England, including the replacement schools, numbered 543 in 2003 and 570 in 2003 – 181 schools in each group in the first year and 190 in the second. There is no replacement of non-responding pupils.

Response in England

Table 1 shows the response rates in England at school and pupil levels. In both years, these fell well below the OECD average. The 'before replacement' school rate (BR) refers to response among initial schools. The 'after replacement' rate (AR) measures response among all schools that are approached, whether an initial school or a replacement school. However, replacements, if approached, are excluded by the survey organisers from the denominator of the AR, which is a cause of some controversy (Sturgis et al. 2006). Their inclusion in the denominator would result in rates in England of only 51 percent in 2000 and 56 percent in 2003 (our calculations). As this reflects, replacement schools were substantially less likely to respond than initial schools.

Table 1 here

Automatic inclusion of a country into the OECD reports depends on the level of response achieved. The PISA Consortium, which is the body overseeing the survey, requires a minimum BR of 85 percent for schools or, where this rate was between 65 and 85 percent, an 'acceptable' level of the AR. (The acceptable level rises by one percentage point for every half point that the BR falls below the 85 percent threshold.) The threshold for pupil response is 80 percent. If a country does not meet these requirements, it is asked to give evidence that its achieved sample of responding pupils in responding schools is representative of the survey population and the Consortium then takes a decision on inclusion. This request was made of England in both 2000 and 2003.

In 2000, school response in England fell far short of the BR threshold and the AR was also well below the acceptable level. After evidence was provided on the characteristics of responding schools, the UK was included into the OECD reports on the 2000 data (e.g. OECD 2001). In 2003, England met neither the school nor the pupil response thresholds. The evidence from the analysis of non-response bias that was provided by the Department for Education and Skills (an analysis undertaken by us) was judged to indicate potential problems at the student level, although the Consortium argued that it was 'not possible to reliably assess the magnitude, or even the direction, of this non-response bias' (OECD 2004: 328). As a consequence, the UK was excluded from the OECD reports on the 2003 data.

We restrict attention in this paper to the 2000 and 2003 PISA data, but we note here that the survey has been repeated in England in 2006 and 2009. In 2006, both the school 'after replacement' response rate and the pupil response rate were reported as 89 percent (Bradshaw et al. 2007: 14-15). At the time of writing, it seems that the 2009 rates also show an encouraging increase.

Auxiliary information

The UK is unusual in having national tests and public exams of children's learning at several ages. These are most developed in England. The resulting data provide a rich source for analysis of student performance and progression, e.g. Chowdry et al. (2008). For us, once linked to the PISA survey data, they allow for respondents and non-respondents to be compared on the basis of assessments taken not long before and shortly after the survey was conducted, together with comparison with population values. This is our principal auxiliary information.

We have access to results from 'Key Stage 3' (KS3) tests in English, maths and science – note the close correspondence with the subjects covered in PISA – taken typically at age 14, and 'Key Stage 4' (KS4) public exams taken at age 15 or, more commonly, at 16. The latter are General Certificate of Secondary Education exams taken in a wide range of subjects and General National Vocational Qualifications, known respectively as GCSEs and GNVQs. We focus on three measures: the average points scored by a child across the three KS3 tests, the total points scored in the KS4 exams, where the higher the grade achieved in any subject the higher the points attributed (there are standard equivalences for GSCEs and GNVQs), and whether the child passed five or more subjects in the KS4 exams at grades A^*–C, a measure that receives a lot of attention in debate in England on school effectiveness. KS3 tests are mandatory within state-funded schools but the information is typically missing for children attending private schools.

Administrative records also provide auxiliary information on the child's gender and whether he or she receives Free School Meals (FSM), a state benefit for low income families. Information on FSM is not available at the pupil level for 2000 although we do know the percentage of individuals receiving the benefit in the child's school.

Critically, the achievement measures have a high correlation with PISA test scores for responding pupils – see Table 2. Figure 1 plots the PISA maths score in 2003 against the KS4 total points measure. Appendix A explains the merger of the PISA data with the administrative records and hence the samples on which these statistics are based. The high correlations mean we have ideal auxiliary variables, i.e. variables that are strongly associated with the survey outcomes of most interest.

Table 2 here

Figure 1 here

National registers and the PISA sample

We have access to the auxiliary information just described for all 15 year olds. This information is contained in the Pupil Level Annual School Census and the National Pupil Database, a combination we refer to as the 'national registers'. We define five groups of 15 year olds to guide our analysis of biases:

i) the PISA survey population of pupils in England schools;
ii) all pupils in schools sampled for PISA;
iii) all pupils in responding schools;
iv) all sampled pupils in responding schools;
v) responding pupils.

The survey population consists of the pupils in the PISA target population of all 15 year olds, less permitted exclusions. (NB all 15 year olds are obliged to be in schools.) In practice, our definition of group (i) departs a little from this. First, there is the issue of whether the national registers include all the target population. For 2003, this seems to be the case. However, the registers for 2000 exclude the small minority of pupils who were not entered for any KS4 public exams. In the 2003 data, this group represents about 2 percent of the cohort. Second, we are unable to apply all the exclusions from the target population that are permitted within PISA. Permitted exclusions of schools are those in remote areas, or with very few eligible pupils, or catering exclusively for non-native English speakers or for pupils students with ‘statemented’ special educational needs (SEN); permitted exclusions of pupils within included schools are children with limited proficiency in English or with statemented SEN. (These are main criteria in 2003; those in 2000 are similar but sometimes formulated differently: OECD 2004: 320-2, OECD 2001a: 232.) In practice exclusions are small, accounting for 5.4 percent of the population in England in 2000 and the same percentage again in 2003 for the UK as a whole (Micklewright and Schnepf 2006: 10). We are able to omit special schools catering for SEN students in both years. In 2003 we can omit all pupils ‘statemented’ with SEN in other schools but are unable to do so in 2000 when the registers lacked the SEN status of individual pupils. Our school and pupil exclusions in 2003 totalled 4.7 percent of pupils in the register, suggesting that we mirrored the main exclusions carried out in practice in PISA in this year.

We define group (ii) to include all sampled schools, initial or replacement, including replacements that were not asked to participate.

Appendix A describes the linking of the data from the national registers with files of sampled PISA schools and pupils and files of respondents in order to create datasets for groups (ii) to (v). We are unable to link 3.8 percent of sampled pupils in 2003 and 6.2 percent in 2000. Observations that are not linked do not enter our analysis and are excluded from all groups. We have linked information for 3,641 respondents in 2003 and 3,923 in 2000 and these samples form our basis for group (v).

Weights

Design weights are needed at both school and pupil levels. Although a self-weighting design is the aim, in practice this is not achieved exactly since actual school size may differ from that indicated in the sampling frame; some schools have less than 35 pupils; exclusions need to be accounted for. Weights are also provided on the OECD PISA website that adjust for non-response (see Micklewright and Schnepf 2006). These incorporate the design weights. The OECD school weights adjust for the level of response in each stratum. Since the strata are constructed on the basis of schools' past KS4 results, the adjustment is based *de facto* on schools' average achievement, thus taking into account the pattern of response as well as the level.

The OECD pupil weights take into account the level of response within any school but not the pattern. In general, the adjustment factor is the ratio of the number of students who were sampled to the number who responded and is therefore the same for all responding pupils. The pupil weight also incorporates the OECD school weight.

Our analysis in Section 5 includes a comparison of the impact of OECD weights with the design weights. This shows the extent to which the OECD's adjustment factors correct for biases induced by the pattern of response. At the school level at least, the OECD weights offer some hope of achieving this. Our own response weights that we compute in Section 4 allow in addition for the pattern of pupil response. The next section shows the pattern of pupil response to be critical for the extent of non-response bias.

3. Biases in estimates of achievement parameters based on auxiliary information

From population to responding sample

Table 3 compares the five groups identified in the previous section with respect to the auxiliary information. We apply the design weights only, since we wish to see the full effect of non-response (and sampling variability) and begin by describing the results for 2003. Compared to the population (i), the responding PISA sample (v), over-represents girls and under-represents children from low-income families (FSM receipt). The differences are easily statistically significant at conventional levels. For gender composition, the largest difference is between groups (ii) and (iii) and groups (iii) and (iv), reflecting school response and pupil sampling respectively. For receipt of FSM, differences arise at all stages. The movement from stage to stage almost always reduces the percentage male and the percentage with FSM.

Table 3 here

What about measures of achievement? The means of both the test score variables for responding pupils are higher than the population values. The percentage changes are very different but in terms of population standard deviations the KS3 variable mean rises by nearly 0.1 units and KS4 mean by about half as much again. These are not trivial changes and are easily statistically significant at conventional levels. There is a slight fall following school response, (ii) to (iii), but otherwise the trend is for the mean to rise, with the main change coming at the last stage following pupil response, (iv) to (v). The standard deviations tend to decline, most obviously for the KS4 variable – a fall of 12 percent – and again the largest change comes with pupil response. The top half of Figure 2 shows the changes in mean and standard deviation for the KS4 score and summarises the key findings: (1) responding pupils have higher average achievement and show less dispersion in scores than the population; (2) this is driven in particular by pupil response; but (3) pupil sampling also appears to be a factor.

Figure 2 here

The next two rows in the table show the implication of the changes in mean and variance for the percentage of each group reaching a given threshold of achievement. The percentage achieving five or more good subject passes at KS4 – a measure commonly used in public debate on pupil achievement – is five points higher in the PISA sample than in the population. The second measure shows the percentage beneath a very low standard – the bottom decile of KS4 points in the population. (The figure is not exactly 10 percent in the population due to the lumpiness in the distribution.) Here the impact of a rise in mean and a fall in variance reinforce each other, and the PISA responding sample shows marked under-representation of pupils at this very low level of performance. By contrast, the percentage in the final sample with scores above the top decile in the population is very close to 10 percent (not shown), the effects of the changes in mean and variance here cancelling out.

The picture for 2000 is broadly similar, at least as far as the achievement variables are concerned (gender composition hardly changes across the groups): there is no indication that the slightly higher response rates in 2000 were associated with higher data quality. The rises in the means between the population and the final sample are rather larger in population standard deviations terms, by 0.13 for KS3 score and 0.20 for KS4 points. (Our inability to remove 'statemented' SEN pupils in normal schools in 2000 from groups (i)-(iii) will have held down the population values a little.) The standard deviations fall by 8 and 9 percent respectively. The percentage of pupils with at least five good KS4 subject passes rises by 7½ points. These differences are strongly statistically significant. The lower half of Figure 2 summarises the changes for the mean and standard deviation of KS4. The most obvious difference from 2003 is that school response is associated with as big an increase in the mean as pupil response.

Pupil response

Table 3 shows that the main source of non-response biases came through pupil response, at least in 2003, and we now investigate this in more detail. Differences between respondents and non-respondents are strongly significant in both years – see Table 4. (The exception is the percentage male in 2003.) The sizes of several of the differences are striking, for example Free School Meals receipt in 2003 (not measured in 2000): receipt among non-respondents is a third higher than among respondents. The KS3 and KS4 points means in 2003 differ by nearly 30 percent and 40 percent respectively of the population standard deviation values. The percentage of pupils with five good KS4 passes is higher for respondents by 17 percentage points in 2003 and by 14 points in 2000. The standard deviation of KS4 points for respondents is 15 percent lower than the value for non-respondents. Given a non-response rate of some 20-25 percent of pupils, these differences are sufficient to generate non-negligible biases – shown in Table 3.

Table 4 here

We build on Table 4 by estimating a logistic regression model for the probability that a sampled pupil responds to PISA. Let $Y_i = 1$ if pupil *i* responds and

$Y_i = 0$ if he or she does not; $\text{prob}(Y_i = 1) = 1/[1+\exp(-\boldsymbol{\beta}\mathbf{X}_i)]$. The model is estimated separately for 2000 and 2003. Estimates of the parameters $\boldsymbol{\beta}$ are given in Table 5.

Our approach to model selection is conservative and the specification of $\mathbf{X}_i$ is simple. We focus on a suitable functional form for the auxiliary information on achievement, where non-linearity was immediately evident. Using the KS4 total points variable, we settled on a piece-wise linear functional form – model 1. We also show the results of a quadratic specification – model 2. We tested for the inclusion of KS3 points but the variable proved insignificant, controlling for the KS4 score. The knots are at about the 13th, 60th, and 97th percentiles of KS4 points in 2003 and at the 12th and 80th percentiles in 2000. The first two estimated coefficients in the piece-wise models and both coefficients in the quadratic models are very well determined. In both years, the probability of response rises substantially with KS4 points and then flattens out. (The turning point for the quadratic models is close to the top of the range of the data.) Figure 3 illustrates the results for 2003. The predicted probability of response rises from about 0.5 at low levels of KS4 points to around 0.8 at high levels.

Table 5 here

Figure 3 here

In 2000, the dummy for boys is significant at the 1 percent level and indicates a *ceteris paribus* increase in the probability of response of about 4 percentage points (evaluating at the mean probability of response), as in the bivariate analysis in Table 3. The probability is about 8 points higher for pupils in schools in the West Midlands. Neither variable has a significant impact in 2003 (we do not include the region dummy in this case). We also exclude from the models two other variables that were insignificant, measuring school type and, notably, receipt of Free School Meals. Controlling KS4 exam scores, we cannot reject the hypothesis that children from low income families have the same probability of responding as other children. The difference in Table 3 merely reflected the association of low income with low academic achievement.

The models in Table 5 do not allow explicitly for school effects. Schools organise the PISA testing of pupils and they may present the survey to their pupils in different ways that affect pupil response. Or there may be peer effects in pupil response. In either case the response probability will vary by school, holding constant individual characteristics. We experimented with simply adding a set of school dummies to the model to pick up such effects. These were not unsuccessful, improving the models' goodness of fit. However, the KS4 coefficients changed little and when we used these extended models to revise the propensity score weights described in the next section, the impact on our estimates of bias changed very little.

4. Construction of new weights

We construct two alternative sets of new weights to adjust for non-response bias. The first set uses the logistic regression models of Table 5 to construct propensity score or inverse probability weights (Little 1986). We use the results of model 1 to calculate a pupil response adjustment factor, equal to the inverse of the predicted probability of response. We then take the OECD weight described earlier and replace its pupil response adjustment factor, which accounts only for the level of response in each school, with our new factor that takes account of the variation of pupil response with cognitive achievement. In this way, the new weighting variable retains the adjustment for design and for the level and pattern of school response in the OECD weight while introducing adjustment for the pattern of pupil response. We refer to the resulting variable as our 'propensity score weight' although it also contains other elements. The new weight does not explicitly adjust for variation in the average level of pupil response across schools that is unrelated to variables included in the logistic regression models; inclusion of school dummies in the models picks this up but, as noted, results with weights based on this richer specification were very similar.

Our second set of weights is based on the generalised regression (GREG) estimator (Kalton and Flores-Cervantes 2003; Särndal and Lundström 2005). These weights are derived from a linear regression model fitted to the survey variables of interest with the auxiliary information as explanatory variables. The resulting estimator may be interpreted as using this regression model for prediction. There is a number of reasons why the weighted estimators arising from the use of GREG weights might be preferred to those from the first approach. These weights exploit the availability of the auxiliary information for the entire population and, as a result, adjust for the impact of response and sampling variability on the achieved sample composition at both school and pupil levels. In terms of our analysis of Section 3, the application of the weights produce mean values of auxiliary variables in group (v) that are equal to those in group (i). The GREG weights may be expected to produce more precise estimates, with the gain in efficiency depending on the predictive power of the linear regression model. The validity of the bias adjustments for both sets of weights depends on (different) modelling assumptions, but the GREG estimator may be expected to be more robust to these assumptions when the predictive power of the auxiliary information is strong. See Särndal and Lundström (2005, sect. 6.1) for further discussion of advantages of the GREG weighted estimator.

We calculate separate GREG weights for each of the three PISA measures of cognitive achievement in maths, science and reading. (It is common to calculate just a single GREG weight in multipurpose surveys but this constraint seems unnecessary for our purposes.) Appendix B reports the results of three regression models estimated for the samples of PISA respondents in each of 2000 and 2003. The dependent variables are the PISA scores. The explanatory variables are the KS3 test and KS4 exam scores and other auxiliary information. The models explain around 70 percent of the variance in the achievement variables. We then use the results, as described in Särndal and Lundström (2005), to construct weights. The models for maths and science in 2003 and reading and science in 2000 have the same specification which implies that the GREG weights for the subjects concerned are identical.

Table 6 gives the correlations between the four sets of weights at our disposal: the design weights, the OECD weights, our propensity score weights and our GREG weights for reading. The correlations are far below 1.00. For example, in 2003 the propensity score weight and the GREG weight both have correlations of less than 0.5 with the OECD weight. We investigated whether outliers could have attenuated these correlations by trimming the weights to between 1/3 and 3 times the mean weight. This led to almost no changes with the 2000 correlation matrix and one or two decreases with the 2003 values. It appears therefore that there are more fundamental reasons for the differences between the weights.

Table 6 here

5. Biases in estimates of achievement parameters based on PISA scores

We now gauge the extent of non-response bias in estimates of achievement parameters that are based on PISA test scores for respondents – of obvious interest for users of the achievement data in the 2000 and 2003 samples. We apply our propensity score weights or our GREG weights when estimating a parameter of interest and then compare the results with those obtained when using the design weights. We also test the use of the OECD weight variable. Table 7 gives results for the mean and the percentage below a score threshold that is emphasised in OECD reports on the survey – students below level 2 are defined as having 'inadequate' or only 'limited' knowledge. Threshold levels were not provided by the survey organisers for science in 2003 or for science or maths in 2000.

Tables 7a and 7b here

Compared to the use of design weights, in both years the application of the OECD weights slightly reduces the means and produces a small increase in the percentage of pupils beneath PISA level 2. Use of either of our propensity score or GREG weights has a much larger impact in the same directions in 2003. The two sets of weights produce very similar results. The pattern is a little different in 2000: use of either set of weights pushes down the mean relative to the value obtained with the design weights and the amount of change is similar to that in 2003 in the case of the propensity score weights. But the change in the mean is much larger when using the GREG weights. This difference between the use of our two alternative sets of weights for 2000 can be understood looking at Figure 2, which shows how KS4 scores change while moving from the population, group (i), to the responding sample of pupils, group (v). The use of the propensity score weight can be expected to correct largely for the bias introduced by the pattern of pupil response – the difference between groups (iv) and (v). But the GREG weights in addition correct for differences between groups (i) and (iv), which, in contrast to 2003, were substantial in 2000 due to the pattern of school response.

Our estimates of the non-response biases are obtained by subtracting the estimates based on our weights from the estimates based on the design weights. The upward bias in the estimates of the mean from the achieved sample of respondents is about 7 to 9 points in 2003. Curiously, the estimated standard errors show that the estimate of bias is better determined when using the propensity score weights but it is still significant at the 5 percent level when using the GREG weights. The downward bias for the percentage below PISA level 2 in 2003 is estimated to be about 3 percentage points for both maths and reading. This reflects both the upward bias for the mean and the downward bias (not shown) for the standard deviation, which we estimate to be about 2 to 3 percent. The estimated bias in the mean is about 0.06 of a standard deviation, which is between one third and two-thirds of the figures estimated for the means of auxiliary variables discussed in Section 3.

The estimates of the extent of the biases for 2000 are not dissimilar on the basis of the propensity score weights but they are substantially larger with GREG weights, especially for the mean. We estimate biases of between 4 to 15 points for the mean and 2 to 4 points for the percentage below PISA level 2 in reading. The figures for biases in the mean are not that well determined when using the propensity score weights – the p-values vary from about 0.07 to 0.02 – and this contrasts with the figures for 2003, but are more precise with the GREG weights. The estimated biases for the percentage below PISA level 2 are well determined, as with our estimates for 2003.

Finally, comparison of the results for design weights and the OECD weights show that the latter do little to correct for the biases we have identified. This reflects the lack of adjustment in the OECD weights for the pattern of pupil response, which we have emphasized to be the principle source of bias.

In addition, we also considered biases in regression parameters that summarise the association between achievement and measures of family background that are recorded in the survey data. This recognises one of the principal lines of research that have been pursued with PISA data, e.g. Schulz (2006) and Woessmann (2004). It is sometimes argued that when non-response biases affect estimates of population means based on data from a survey, estimates of covariances may be free of bias. Table 8 shows results from a simple bivariate OLS regression of the PISA maths score on a continuous measure of parental occupational status, the index proposed by Ganzeboom et al. (1992) that is one of the main measures of family socio-economic background recorded in the PISA database. The regression model is estimated in turn with each of the four weight variables used in Tables 7a and 7b. Besides constant and slope coefficient we show the estimates of the covariance between the two variables and the variance of the occupation index – the OLS slope coefficient is the ratio of these two figures. When the propensity score weights are used, the slope coefficient is about 9 percent larger than when the design weights are used and the change is equivalent to about one standard error of the parameter estimate. The last two columns in the table show this to be driven by the change in the covariance of the two variables rather than by the change in the variance of the occupational index. However, the change is much smaller when we apply the GREG weight. As a result, it is hard to draw any general conclusions from these results other than it cannot be assumed that estimates of regression coefficients will be free of bias from non-response.

Table 8 here

6. Discussion of Biases

How large are the biases we have estimated? One way of judging this is to consider the contribution of bias to 'total survey error', which combines sampling and non-sampling errors in the estimate of a parameter. This is conventionally measured by mean squared error (MSE) defined as the square of the bias plus the square of the standard error. Biases can arise for various reasons but we restrict attention to the pupil non-response biases that we have been able to estimate. The quadratic terms in the formula for MSE implies that as bias rises above the standard error, it will quickly come to dominate. Where the bias is less than the standard error, most of MSE will be due to sampling variation.

Our estimates of the biases are considerably larger than the estimated standard errors of the parameters concerned. In the case of the auxiliary variable means, the estimated biases shown in Table 3 produced by pupil response, the main source of bias in 2003, represents over 90 percent of MSE. Likewise, in the case of the PISA test scores, estimated bias of 7 to 9 points in the mean may be compared with estimates for the standard error of the mean of about 2 to 4 points. Again, bias dominates MSE. We estimate bias in the standard deviations of 2 to 3 points (not shown in Tables 7a and 7b) compared with estimates for the standard errors of the standard deviations of 1½ to 2 points. (The standard errors for 2003 are taken from an Excel file of results for England available on the OECD PISA website; figures for 2000 are given in Gill et al. 2002.) Only in the case of the regression parameter estimates in Table 8 do we obtain an estimate of bias that is equal to or less than one standard error.

Viewed in this way, relative to the impact of sampling variation, the estimated biases are, in general, large. This is not uncommon in large surveys: the larger the survey sample the smaller the standard error and hence bias comes to dominate. However, in sub-samples, e.g. children from particular socio-economic backgrounds or types of schools in the case of PISA, sampling variation may come to be more important since, *ceteris paribus*, standard errors rise as sample size falls, while bias could rise, fall or stay the same. We suspect that the PISA Consortium's decision to exclude the UK from OECD reports on the 2003 data was driven by this view of the likely contribution of bias to total survey error. Commenting on the minimum thresholds set for acceptable levels of response, for example 80 percent for pupils, it was noted:

> 'In the case of countries meeting these standards, it was likely that any bias resulting from non-response would be negligible, i.e. smaller than the sampling error' (OECD 2004: 325).

However, as we have seen, in practice bias can still exceed sampling error when the threshold is met. Pupil response in England in 2000 met the required level but the biases we have estimated for this year are not surprisingly about as large as those in 2003 when response was only a little lower, and even larger in the case of the mean when we use the GREG weights. The situation in England makes one wonder about the extent of biases in countries with response rates not far above the threshold. Australia, Austria, Canada, Ireland, Poland and the US all had pupil response rates of between 82 and 84 percent in 2003 (OECD 2004: 327).

Another way to consider the size of the biases is to check their impact on the picture shown by PISA of differences in learning achievement between countries. We calculated how many places England would move in a 'league table' of 2003 rankings of countries by their mean scores if the English means for reading, maths and science were adjusted downwards by the estimated bias of 7 to 9 points. (We consider all countries participating in the survey in that year, including those not in the OECD.) England shifts by 3 places for maths, 2 for science, and none for reading. Likewise, for the percentage of pupils below PISA level 2, England would move by 3 places for both maths and reading. Viewed in this way, the effect of the biases appear more modest.

7. Conclusions

We have investigated non-response biases in two rounds of PISA in England: in 2000, when response rates were deemed high enough for OECD to publish the results, and in 2003, when response rates were a little lower and deemed of sufficient concern for the results not to be published. We have found clear evidence of biases, but there is no indication that the slightly higher response rates in 2000 were associated with higher data quality. Indeed there is some evidence that the (absolute) biases in the mean achievement scores are greater in 2000 than 2003. This underlines the danger of using response rate thresholds as a guide to data quality. The higher response rates in PISA in England in 2006 and (it seems) 2009 are encouraging, but should not be treated as definitive evidence of higher data quality

We have considered a number of alternative weighting methods to adjust for non-response bias when estimating the distribution of different measures of achievement. We have found that very little of the bias is removed by weighting methods, such as those provided by OECD, which only allow for differences in (school or pupil level) sampling probabilities, for school-level non-response or for differences in overall pupil response rates within schools. The most important source of bias seems to be associated with within-school differences in response by different kinds of pupils. We have shown how to adjust for such bias using auxiliary data on the results of national tests of achievement, which is available at the population level and is linked to the pupil-level survey data. Our preferred weighting approach employs the generalized regression (GREG) estimator, which demonstrates considerable gains in precision compared to the other weighting methods, as a result of the strong correlations between the survey achievement measures and the auxiliary tests. The sizes of the bias-adjustments can be considerably larger than the estimated standard errors of the parameters concerned, as discussed in the previous section.

Appendix A: Linking PISA survey data to the national registers

Our linking of the survey and register data is not perfect. There are a (very) few schools that were sampled for PISA for which we could find no record in the national registers. Within schools that are successfully linked, we can find no record in the registers for some sampled pupils, especially in 2000 when the register data exclude pupils with no KS4 entries. In total, as a result of either cause, we are unable to link 3.8 percent of sampled pupils in 2003 and 6.2 percent in 2000. Schools and pupils that are not linked are excluded from all groups (i) to (v). In the case of responding pupils whom we are unable to link, we can compare their characteristics recorded in the survey data with those of linked pupils. In 2003, the mean PISA achievement scores are slightly higher for pupils who are not linked but in each case – maths, science and reading – the difference is not significant at the 10 percent level. In 2000, the pupils who are not linked have considerably lower mean scores (by about 20 points), consistent with the register data excluding pupils with no KS4 entry, and the differences are statistically significant at conventional levels. All our results for PISA variables in the paper were obtained with observations that we could link to the national registers and in the case of respondents this may account for any slight differences from results for England published by the survey organisers.

	2000			2003		
	original number	linked number	% loss	original number	linked number	% loss
Approached schools	306	302	1.3	276	273	1.1
Responding schools	155	152	1.9	159	157	1.3
Non-responding schools	151	150	0.7	117	116	0.8
Sampled pupils	5,164	4,846	6.2	5,213	5,015	3.8
Responding pupils	4,120	3,923	4.8	3,766	3,641	3.3
Non-responding pupils	1,044	923	11.6	1,447	1,374	5.0

Notes: There are 122 non-responding schools in the data file we received for 2003. However, five of these are special schools. Under the assumption that they were wrongly approached, we exclude those schools from our analysis. The sampled pupils in the table exclude pupils ‘statemented’ with SEN and pupils in schools with pupil response below 25 percent, which are treated in PISA as non-responding schools.

Appendix B: Regression models underlying the GREG weights

The table reports least squares estimates of the coefficients of the regression models described in Section 4. The explanatory variables were chosen using forward selection. In general this gave the same result as backward selection.

	2003			2000		
	reading	maths	science	reading	maths	science
Male	-13.31	22.19	19.68	-10.03	25.15	13.96
	(1.69)	(1.67)	(1.86)	(1.68)	(1.99)	(2.16)
KS3 average score	6.28	7.51	7.47	6.09	5.57	5.80
	(0.24)	(0.23)	(0.26)	(0.25)	(0.30)	(0.32)
KS3 missing	16.68	20.15	16.03	17.55		21.74
	(4.63)	(4.55)	(5.07)	(4.79)		(5.99)
KS4 5+ good grades (dummy)	7.39					
	(2.87)					
KS4 nos. of good grades		1.58	2.13		-1.90	
		(0.58)	(0.65)		(0.70)	
KS4 average points score	12.96	12.65	13.84	10.69	16.68	7.33
	(2.06)	(2.06)	(2.29)	(2.40)	(2.36)	(2.96)
KS4 capped points score (best 8 subjects)	1.44	0.79	1.31	1.10		1.50
	(0.32)	(0.32)	(0.36)	(0.41)		(0.52)
KS4 total points score	-0.44	-0.34	-0.61	0.85	1.32	0.67
	(0.17)	(0.20)	(0.23)	(0.24)	(0.20)	(0.31)
Free School Meals (FSM) variable missing	24.30					
	(5.15)					
Proportion of pupils with FSM in the school				-40.03	-54.43	-67.09
				(7.09)	(8.54)	(9.09)
Private school		29.77	27.11	19.44	26.69	16.09
		(5.08)	(5.66)	(5.14)	(4.50)	(6.33)
Constant	192.52	151.04	150.01	196.42	208.26	216.65
	(5.88)	(6.15)	(6.85)	(6.20)	(8.38)	(7.98)
Observations	3,641	3,641	3,641	3,923	2,181	2,177
R-squared	0.68	0.70	0.68	0.71	0.72	0.71

Note: Estimated standard errors in parentheses. The dependent variables are the averages of the five ‘plausible values’ for achievement in each subject that are provided by the PISA organizers for each individual. These are random draws from an estimated ability distribution for individuals with similar test answers and backgrounds. The sample sizes are lower for maths and science in 2000 as tests in these subjects were conducted for a sub-set of pupils in this year.

References

Adams R (2003) 'Response to "Cautions on OECD's Recent Educational Survey (PISA)"' *Oxford Review of Education* 29(3): 377-89.

Bradshaw J, Sturman L, Vappula H, Ager R, and Wheater R. (2007). *Achievement of 15-year-olds in England: PISA 2006 National Report* (OECD Programme for International Student Assessment). Slough: NFER.

Chowdry H, Crawford C, Dearden L, Goodman A, and Vignoles A (2008) *Widening Participation in Higher Education: analysis using linked administrative data*, Institute for Fiscal Studies. http://www.ifs.org.uk/publications/4234

DfES (2005), 'PISA 2003: Sample design, response and weighting for the England sample', unpublished document.

Ganzeboom H, De Graaf P, and Treiman D J, with De Leeuw J. (1992) 'A standard international socio-economic index of occupational status', *Social Science Research*, 21, 1–56.

Gill B, Dunn M and Goddard E (2002), *Student Achievement in England*, London: The Stationary Office.

Groves R (1989) *Survey Errors and Survey Cost*, New York: John Wiley.

Groves R (2006) 'Nonresponse rates and nonresponse bias in household surveys' *Public Opinion Quarterly*, 70(5): 646–75.

Kalton G and Flores-Cervantes I (2003) 'Weighting methods' *Journal of Official Statistics* 19(2): 81-97.

Little, R.J.A. (1986) 'Survey nonresponse adjustments for estimates of means' *International Statistical Review*, 54: 139-157.

Lynn P (2004) 'The use of substitution in surveys', *The Survey Statistician*, 49: 14-16.

Micklewright J and Schnepf S V (2006) *Response bias in England in PISA 2000 and 2003*, DfES Research Report 771.

OECD (2001), *Knowledge and Skills for Life. First Results from PISA 2000*, Paris: OECD.

OECD (2004), *Learning for Tomorrow's World. First Results From PISA 2003*, Paris: OECD.

OECD (2005), *PISA 2003 Technical Report*, Paris: OECD.

Prais S G (2003) 'Cautions on OECD's Recent Educational Survey (PISA)' *Oxford Review of Education* 29 (2): 139-63.

Särndal, C-E. and Lundström, S. (2005) *Estimation in Surveys with Nonresponse*. Chichester: Wiley.

Schulz, W. (2006), 'Measuring the socio-economic background of students and its effect on achievement in PISA 2000 and PISA 2003', Paper presented at the Annual Meetings of the American Educational Research Association (AERA), San Francisco.

Sturgis P, Smith P and Hughes G (2006) *A study of suitable methods for raising response rates in school surveys*, DfES Research Report 721.

Vehovar V (1999) 'Field substitution and unit nonresponse', *Journal of Official Statistics*, 15(2): 335-50 .

Woessmann L (2004) 'How equal are educational opportunities? Family background and student achievement in Europe and the United States' IZA Discussion Paper 1284.

Table 1: Response rates in PISA at school and student levels in 2000 and 2003 (%)

	England		OECD average	
	2000	2003	2000	2003
School 'before replacement'	59	64	86	90
School 'after replacement'	82	77	92	95
Pupil	81	77	90	90

Source: Response rates for OECD countries from OECD (2001) and OECD (2004); figures in table are simple averages of the country values (including the UK); response rates for England from Gill et al. (2002) and DfES (2005).

Table 2: Correlations between achievement measures based on PISA test scores and on auxiliary information

a) 2000

	KS3 avg. pts.	KS4 tot. pts.	PISA reading	PISA maths	PISA science
KS3 average points	1.00				
KS4 total points	0.83	1.00			
PISA reading	0.82	0.80	1.00		
PISA maths	0.82	0.78	0.91	1.00	
PISA science	0.82	0.78	0.94	0.93	1.00

b) 2003

	KS3 avg. pts.	KS4 tot. pts.	PISA reading	PISA maths	PISA science
KS3 average points	1.00				
KS4 total points	0.82	1.00			
PISA reading	0.80	0.74	1.00		
PISA maths	0.82	0.72	0.90	1.00	
PISA science	0.81	0.72	0.93	0.94	1.00

Notes: Correlations are computed for unweighted data. KS3 scores are missing for 11 percent of the PISA respondents in 2000 and 8 percent in 2003, which is largely explained by the KS3 tests not being taken in most private schools. The PISA points scores are averages of the five 'plausible values' estimated by the survey organizers for each individual.

Table 3: Estimates of characteristics of pupils using auxiliary information

	Popl. (i)	Sampl. schools (ii)	Respnd. schools (iii)	Sampl. pupils (iv)	Respnd. pupils (v)
2003					
Male (%)	50.02	49.28	47.48	46.31	46.31
Free School Meals (%)	13.78	12.54	11.89	11.23	10.27
means					
KS3 average points	34.16	34.32	34.18	34.26	34.78
KS4 total points	42.86	43.00	42.55	43.57	45.84
standard deviations					
KS3 average points	6.62	6.63	6.49	6.44	6.29
KS4 total points	21.09	20.74	20.65	19.71	18.51
thresholds					
KS4 5+ good grades (%)	55.79	56.10	55.19	56.45	61.07
< popl. bottom decile KS4 pts. (%)	10.2	9.7	9.7	7.1	4.2
2000					
Male (%)	50.35	50.15	49.50	49.01	49.77
means					
KS3 average points	32.96	32.80	33.30	33.53	33.83
KS4 total points	41.10	41.16	42.46	43.47	44.84
standard deviations					
KS3 average points	6.54	6.46	6.41	6.21	6.03
KS4 total points	19.04	19.01	18.90	18.46	17.34
thresholds					
KS4 5+ good grades (%)	52.10	52.40	54.70	57.02	59.77
< popl. bottom decile KS4 pts. (%)	10.3	10.4	8.9	7.2	4.6

Note: School design weights are applied for groups (ii) and (iii) and pupil design weights are applied for groups (iv) and (v). KS3 points are missing for 8.6 percent of the population in both years and for 7.8 percent of sampled pupils in 2000 and 5.7 percent in 2003. They are typically missing for pupils in private schools.

Table 4: Differences in characteristics between samples of responding and non-responding pupils

Variable	Respondent: Yes	Respondent: No	Difference (Yes-No)	p-value
2003				
Male (%)	46.31	46.33	-0.02	0.99
Free School Meals (%)	10.27	13.73	-3.46	0.00
KS3 average points (mean)	34.78	32.88	1.90	0.00
KS4 total points (mean)	45.84	37.55	8.29	0.00
KS4 5+ good grades (%)	61.07	44.20	16.87	0.00
% below bottom decile KS4 points	4.18	14.84	-10.67	0.00
KS3 average points (SD)	6.29	6.63	-0.33	0.02
KS4 total points (SD)	18.51	21.46	-2.95	0.00
2000				
Male (%)	49.77	45.79	3.99	0.07
KS3 average points (mean)	33.83	32.23	1.60	0.00
KS4 total points (mean)	44.84	37.66	7.17	0.00
KS4 5+ good grades (%)	59.77	45.33	14.44	0.00
% below bottom decile KS4 points	4.63	18.20	-13.57	0.00
KS3 average points (SD)	6.03	6.78	-0.75	0.00
KS4 total points (SD)	17.34	21.69	-4.36	0.00

Note: Design weights are applied. The clustering in the survey design is taken into account when estimating standard errors. In 2003 there are 3,641 respondents and 1,374 non-respondents (3,442 and 1,302 for Free School Meals and 3,423 and 1,304 for the KS3 measure). In 2000, these figures are 3,923 and 923 and, for the KS3 measure, 3,613 and 853 (we do not have information on individual Free School Meals receipt for this year).

Table 5: Logistic regression models of pupil response – parameter estimates

	2000		2003	
	Model 1	Model 2	Model 1	Model 2
KS4 points (0 to 20)	0.104		0.065	
	(0.012)		(0.010)	
KS4 points (20 to 60)	0.016			
	(0.004)			
KS4 points (60+)	-0.030			
	(0.011)			
KS4 points (20 to 50)			0.026	
			(0.004)	
KS4 points (50 to 80)			-0.007	
			(0.005)	
KS4 points (80+)			0.054	
			(0.034)	
KS4 points		0.087		0.060
		(0.008)		(0.006)
KS4 points squared/100		-0.081		-0.047
		(0.010)		(0.007)
Male	0.270	0.268	0.120	0.125
	(0.090)	(0.090)	(0.076)	(0.076)
West Midlands	0.474	0.466		
	(0.164)	(0.162)		
Constant	-0.965	-0.637	-0.747	-0.591
	(0.210)	(0.170)	(0.158)	(0.122)
Observations	4,846	4,846	5,015	5,015

Notes: The mean of the dependent variable is 0.810 for 2000 and 0.726 for 2003. Standard errors are given in brackets and are estimated allowing for clustering of pupils within schools. The first six variables refer to piece-wise linear splines of KS4 points.

Table 6: Correlation of weights: respondents in 2003 and 2000

	Design	OECD	Prop.-score	GREG
2003				
Design	1.00			
OECD	0.61	1.00		
Propensity score	0.39	0.43	1.00	
GREG (reading)	0.49	0.32	0.67	1.00
2000				
Design	1.00			
OECD	0.50	1.00		
Propensity score	0.84	0.56	1.00	
GREG (reading)	0.17	0.19	0.40	1.00

Table 7a: Estimates of characteristics of distribution of PISA test scores using different weights, 2003

Weight	Maths	s.e.	Reading	s.e.	Science	s.e.
Mean						
Design	507.8	3.89	507.3	3.90	520.2	4.10
OECD	506.8	4.14	506.1	4.14	519.0	4.40
Propensity score	501.0	4.39	500.1	4.43	512.8	4.64
GREG	500.4	1.61	498.1	1.65	511.6	1.74
% < PISA level 2						
Design	17.75	1.14	14.65	0.99	n.a.	n.a.
OECD	18.24	1.22	15.16	1.06	n.a.	n.a.
Propensity	20.89	1.34	17.46	1.19	n.a.	n.a.
GREG	20.70	0.77	17.70	0.71	n.a.	n.a.
Differences between means						
Design − P-score	6.8	0.91	7.2	0.91	7.4	0.96
Design – GREG	7.4	3.32	9.2	3.24	8.6	3.50
Differences between % < level 2						
Design − P-score	-3.14	0.34	-2.81	0.31	n.a.	n.a.
Design – GREG	-2.95	0.85	-3.05	0.69	n.a.	n.a.

Notes: Estimates of standard errors of the mean and the percentage below PISA level 2 are calculated separately for each plausible value, taking into account clustering of pupils within schools, and then averaged. For the differences between estimates of the percentages below PISA level 2, the standard errors are estimated by using a single figure for the percentage calculated using the mean of the five plausible values for each pupil and the mean of the thresholds supplied by the survey organizers for each plausible value. Threshold levels were not provided by the survey organisers for science in 2003.

Table 7b: Estimates of characteristics of distribution of PISA test scores using different weights, 2000

Weight	Maths	s.e.	Reading	s.e.	Science	s.e.
Mean						
Design	531.3	4.02	525.7	4.18	535.8	4.37
OECD	531.0	4.41	525.0	4.70	535.3	4.84
Propensity score	527.2	5.20	520.9	5.51	531.0	5.37
GREG	516.8	1.59	510.5	1.59	521.3	1.76
% < PISA level 2						
Design	n.a.	n.a.	11.95	0.91	n.a.	n.a.
OECD	n.a.	n.a.	12.43	1.06	n.a.	n.a.
Propensity	n.a.	n.a.	14.18	1.23	n.a.	n.a.
GREG	n.a.	n.a.	15.68	0.72	n.a.	n.a.
Differences between means						
Design − P-score	4.1	2.22	4.8	2.45	4.8	2.02
Design – GREG	14.5	3.83	15.2	3.88	14.5	4.01
Differences between % < level 2						
Design − P-score	n.a.	n.a.	-2.23	0.49	n.a.	n.a.
Design – GREG	n.a.	n.a.	-3.73	0.71	n.a.	n.a.

`

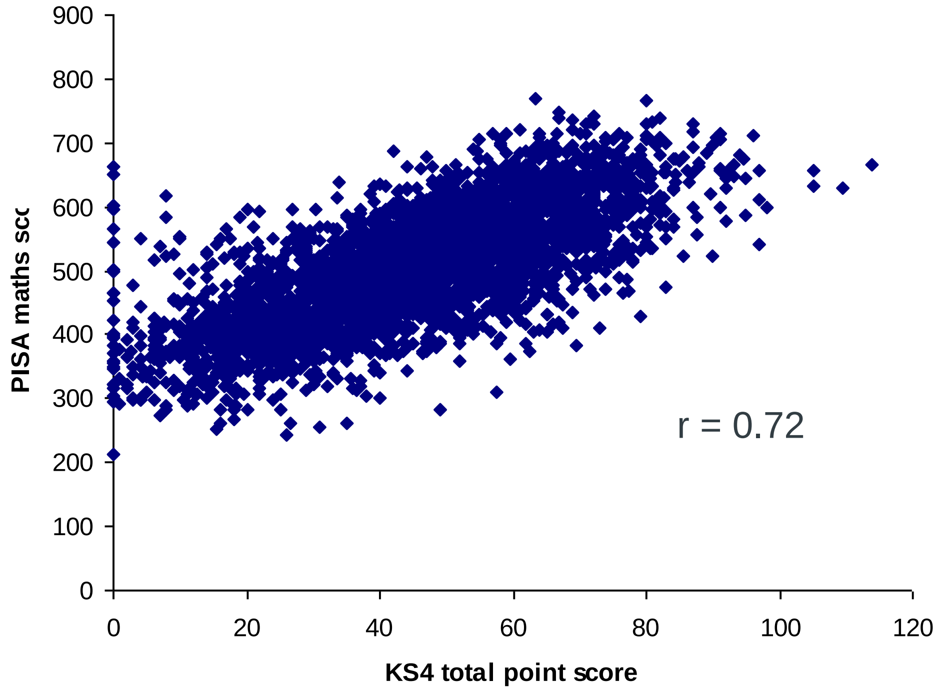
Notes: See Table 7a. Threshold levels were not provided by the survey organisers for maths or science in 2000.

Table 8: Parameter estimates from OLS regression of PISA maths score on index of parental occupation, 2003

weight	constant	slope coefficient	covariance (maths, occo.)	variance (occo.)
Design	413.59	1.95	529.97	271.69
OECD	411.60	1.97	541.27	274.53
Propensity	399.93	2.12	584.93	276.22
GREG	405.57	2.00	541.66	271.32

Notes: The dependent variable is the average of the five plausible values of the PISA maths score; the explanatory variable is an index of parental occupational status (the higher of values recorded for father and mother). The estimated standard error on the slope coefficient varies between 0.12 and 0.14 (clustering of pupils within schools is allowed for). The sample size, allowing for missing values of parental occupation, is 3,394 observations.

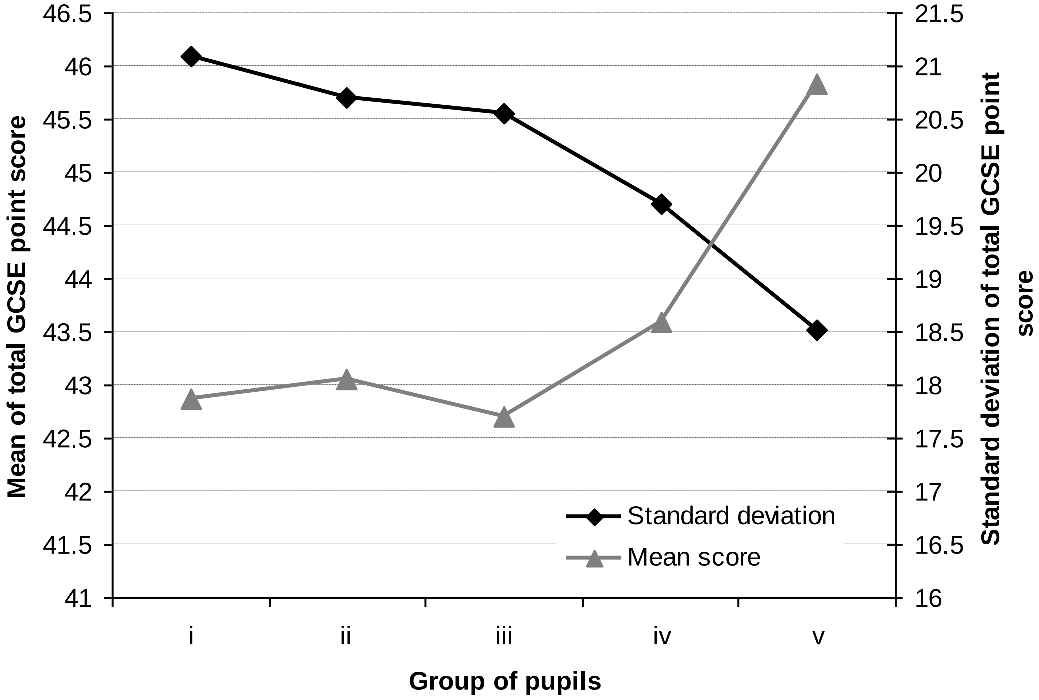
Figure 1: PISA maths score and KS4 total points score: responding pupils, 2003



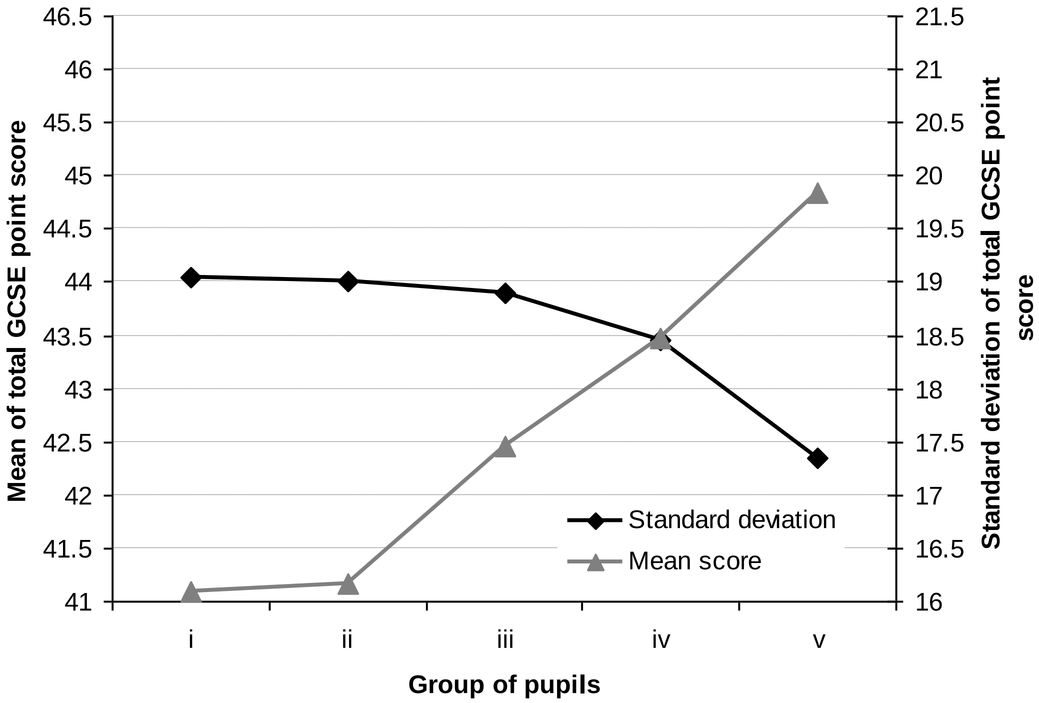
Note: the sample used is responding pupils for whom auxiliary information could be linked – see Appendix A. The PISA maths points score is the average of the five plausible values estimated by the survey organizers for each individual.

Figure 2: Mean and standard deviation of KS4 total point score

2003

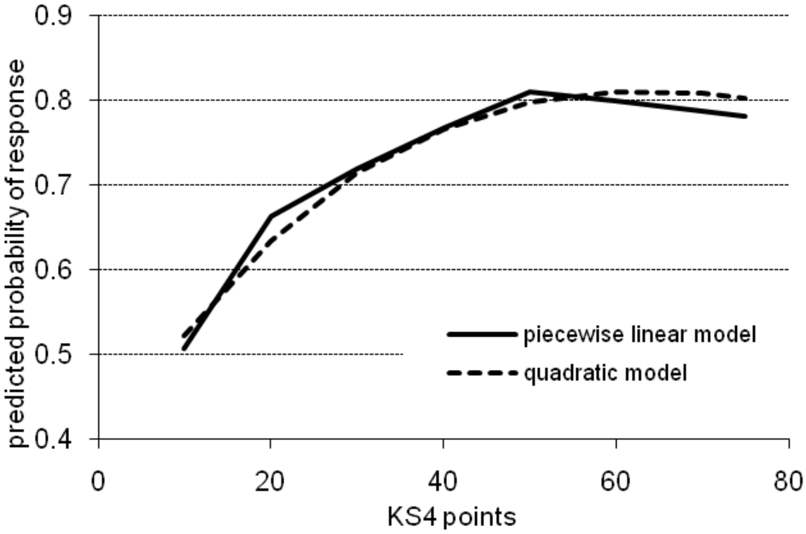


2000



Note: School design weights are used for groups (ii) and (iii) and pupil design weights for groups (iv) and (v). The groups are defined in Table 3 and in the text.

Figure 3: Predicted probability of pupil response by KS4 point score, 2003



Note: The graph shows the predicted probability of response for a boy for KS4 points scores between the 5th and 95th percentiles of the sample based on the models for 2003 in Table 5.