MISUSE OR MISREPRESENTATION OF OECD DATA BY GOVERNMENT MINISTER

The TES reported the schools minister, Sam Gyimah, HAD claimed at the Conservative Conference that 'when the coalition government came to power in 2010 figures from the influential Programme for International Student Assessment (Pisa) tests showed that more than a third of young people were leaving school "unable to read, write or do maths"'.

'The statistics, he said, meant the UK had some of the most "illiterate and innumerate young people in the developed world".'

http://news.tes.co.uk/b/news/2014/09/30/conservative-minister-condemned-after-claiming-uk-students-39-illiterate-and-innumerate-39.aspx

It is not the first time school ministers have misused PISA data.   In October 2012, Sir Andrew Dilnot criticised the Government's use of PISA figures in a letter to David Miliband. In particular, the Government had ignored an OECD warning not to compare the PISA 2009 results for the UK with PISA UK data from the year 2000 because the latter had been found to be flawed. The Government ignored this warning.

Mr Gyimah is wrong when he said PISA proved more than a third of young people were leaving school 'unable to read, write or do maths'.   In reading, the 2009 PISA results showed that 'In England, one per cent of pupils scored below PISA level 1b, compared with the OECD average of 1.1 per cent. At level 1a or below, England had 18.4 per cent, compared with the OECD average of 18.8 per cent.'   (Page 21 http://www.nfer.ac.uk/publications/NPDZ01/NPDZ01.pdf). In maths, the results showed that 'in England, 6.1% of pupils scored below PISA level 1, which was less than the OECD average of eight per cent… The OECD average for the proportion of pupils at level 1 or below, was 22 per cent. England has 19.8 per cent of pupils at these levels.' (Page 28 *op cit*).

Being at Level 1 in PISA tests does not imply a complete inability to read or do maths. Even if it were, 18.4% (reading) and 19.8% (maths) is not 'more than a third' - unless, of course, the minister is being disingenuous and adding 18.4 to 19.8 to reach a total of 38.2. In any case, the proportion attaining Level 1 or below was slightly lower than the OECD average – something the minister appears to have missed.

Perhaps Mr Gyimah was not referring to PISA 2010 but the results of the OECD Adult Skills Survey published last year. These showed English and Northern Irish young people doing badly in a survey of just 23 OECD countries (although they weren't at the bottom). 23 countries do not comprise the entire 'developed world'.   But the OECD issued a warning: the results should be used with caution because of sampling issues.

These sampling issues were caused by non-response by most of the countries including England and Northern Ireland. The OECD asked these countries to do non-response bias analyses but England and Northern Ireland were the only two countries that did not do all of them. This casts doubt on the validity of the test results.

I queried the data with the OECD which stood by the results. The reply by William Thorn of the OECD is included as an appendix. However, the OECD had been concerned enough to warn that the

data should be used with caution. As Sir Andrew said in his letter to David Miliband re the Government's earlier misuse of PISA data:

'The principles of good statistical communication and statistical literacy emphasise the importance of recognising statistical uncertainty and quality weaknesses in the commentary and advice provided to the user. These uncertainties and weaknesses are not just a technical footnote; they are themselves an important part of the evidence, and affect interpretation and meaning. League tables and the presentation of international rankings can be statistically problematic, and require clear and careful commentary alongside them.'

'Clear and careful commentary' was missing from the minister's claim.

I should be grateful if you would investigate the issue of misuse or misrepresentation of OECD data by ministers and issue your comments.

APPENDIX   Copy of email sent to me on 17 October 2013

Dear Ms Downs

Thank you for your interest in the Adult Skills Survey.

You ask whether the survey results for England/Northern Ireland should have been withheld as the level of non-response bias were unknown. Despite the fact that certain elements of the non-response bias analysis were not completed, the overall conclusion was that the data from England and Northern Ireland was of sufficient quality to be published. I have attached the data adjudication report concerning the UK for your information. This can be found in the Technical Report for the Survey.

 Yours sincerely

 William Thorn

William Thorn
Senior Analyst
Directorate for Education and Skills, Programme for the

International Assessment of Adult Competencies

2, rue André Pascal - 75775 Paris Cedex 16
Tel: +33 1 45 24 78 04 – Mob: +33 6 27 95 47 20

william.thorn@oecd.org || www.oecd.org

The United Kingdom (England and Northern Ireland)

Sampling

The United Kingdom followed the PIAAC Technical Standards and Guidelines (TSG) related to sampling and weighting.

· Sampling plan: No issues

· Sample selection

o Home office: No issues

o In field: The theoretical person base weights (THEOR_PBWT) were derived from imputed values of the number of eligible people in the sampled household (NUM_ELG) for 52 cases (49 in England and 3 in Northern Ireland) due to a technical problem with the contact data that the interviewers entered.

· Sample weighting: The Consortium followed the procedures in the PIAAC Weighting and Variance Estimation Plan to create two sets of weights separately for England and Northern Ireland. The United Kingdom (England and Northern Ireland) did not collect age and gender for all sampled persons during the screener. A special adjustment was implemented so that literacy-related nonrespondents with age and gender successfully collected represented those with age or gender not successfully collected.

· Sampling error: The design effect due to unequal weights is 1.35 for England for a sample size of 5 131; and 1.54 for Northern Ireland for a sample size of 3 761. The effective sample size, which is the sample size needed to achieve the same sampling variance as a simple random sample, is 2 425 for England and 2 773 for Northern Ireland. The effective sample size was computed as the number of cases with plausible values divided by the overall design effect (using the literacy component first plausible value), which accounts for both unequal weights and clustering. The United Kingdom's address sample was an equal probability sample in both England and Northern Ireland. Variation in the selection probabilities was introduced from (a) subsampling households for addresses containing multiple households, and (b) the within-household selection at the person level. Further variation in the weights was added through nonresponse and calibration adjustments, although the Consortium followed standard procedures to balance bias and variance.

Coverage and nonresponse bias

·   Population Coverage

o Frame: The combined estimated percentage of the target population excluded from the frame in England and Northern Ireland was 2.0% (individuals living in private residences that are not listed on the "residential" version of the Postal Address File; or, in Northern Ireland, not listed on the NI(POINTER) database).

o Data collection: Not applicable

·   Weighted response rate: 59% for England; 65% for Northern Ireland

·   Nonresponse bias analysis

o Basic: The United Kingdom performed all required analyses and used all required analysis variables (neighborhood characteristics).

- England: The screener response rate varied by region (from 77% in London to 89% in North East England). The highest category screener response rate was 90% (Output Area Classification: terraced blue collar neighborhoods) and the lowest was 74% (Output Area Classification: transient communities). The highest category BQ response rate was 84% (third quintile category of % Indian) and the lowest was 59% (Output Area Classification: transient communities). The lowest regional BQ response rate was in London (61%). Screened households differed from nonscreened households in terms of neighborhood profile. Neighborhoods with a high proportion of residents not born in the UK or of Black or Bangladeshi descent were underrepresented in the screened household sample. London was also underrepresented. Neighborhoods with a high proportion of Black residents were underrepresented in the BQ respondent sample while neighborhoods with a high proportion of older people (aged 65+) and of those with a caring responsibility were overrepresented. London was underrepresented. The classification tree found that region was the only significant screener response rate predictor. The classification tree identified the proportion aged 65+ as the only significant BQ response rate predictor. BQ response rates tended to be higher in neighborhoods with an older-than-average age profile.

- Northern Ireland: The highest category screener response rate was 90% (Output Area Classification: senior communities) and the lowest was 25% (Output Area Classification: Asian communities). Excluding this very small sample size category, the lowest was 70% (Output Area Classification: public housing). It is notable that the screener response rate in the capital Belfast was only 72%. The highest category BQ response rate - excluding categories with small sample sizes - was 86% (Output Area Classification: young families in terraced homes and the lowest was 65% (Output Area Classification: village life). The lowest regional BQ response rate was in the North (75%). Neighborhoods in Belfast were the most underrepresented in the screened household sample. There were no significant profile differences between BQ responders and nonresponders in Northern Ireland. The classification tree identified region and the proportion aged 65+ as strong discriminators of screener response rates. The classification tree identified the proportion aged 65+ as the strongest predictor of BQ response rates.

o Extended: The United Kingdom did not perform all required analyses. Although some paradata were collected, the three agencies responsible for fieldwork did not collect them in a consistent fashion so that it could be used for analytical purpose.

- Analysis 1 – Comparisons of estimates before and after weighting: In England and Northern Ireland, at both the screener and BQ levels, bias in region was reduced through the weighting process as it was used in weighting adjustments. The base-weighted respondent profile was very similar to the base-weighted sampled person profile.

- Analysis 2 – Comparisons of estimates to external totals: In both England and Northern Ireland, large differences were found between PIAAC estimates (using final weights) and Census 2011 totals of employment status, ethnic group and general health. In Northern Ireland, nontrivial differences were also found for qualification. (In England, Census 2011 data on qualification will not be released until late August, 2013.) For age, a large difference was found for category 60-65 (higher in PIAAC), which came as a surprise given the PIAAC control totals were based on census totals updated by birth, death, and immigration/emigration data. The United Kingdom noted that disparities between the PIAAC estimates and Census 2011 totals may reflect the difference in the interview mode (interviewer-assisted vs. self-administered).

- Analysis 3 – Correlation of auxiliary variables and proficiency estimates:

§ England: The correlation between the BQ nonresponse cells and literacy scores was below average at 0.32 (0.35 for numeracy). The correlation between the raking dimensions and literacy scores was average at 0.48 (0.51 for numeracy). The correlation between literacy scores and the combination of nonresponse adjustment cells and raking dimensions was 0.52 (0.56 for numeracy), which was about the average across countries. Although the response rate was 59%, this analysis shows an effective reduction in potential NRB due to the moderate correlation between the survey outcomes and the weighting variables. However, data users need to be cautioned that the analysis is based on correlations between the responding sample (59% of the selected sample) and the weighting variables. That is, the analysis assumes that the same correlations exist for the remaining sampled cases that have no scores (41% of sampled cases).

§ Northern Ireland: The correlation between the BQ nonresponse cells and literacy scores was below average at 0.33 (0.36 for numeracy). The correlation between the raking dimensions and literacy scores was above average at 0.55 (0.58 for numeracy). The correlation between literacy scores and the combination of nonresponse adjustment cells and raking dimensions was 0.57 (0.60 for numeracy), which was higher than average across countries. Although the response rate was 65%, this analysis shows an effective reduction in potential NRB due to the high correlation between the survey outcomes and the weighting variables. However, data users need to be cautioned that the analysis is based on correlations between the responding sample (65% of the selected sample) and the weighting variables. That is, the analysis assumes that the same correlations exist for the remaining sampled cases that have no scores (35% of sampled cases).

- Analysis 4 – Comparisons of estimates from alternative weightings: This analysis was not performed.

- Analysis 5 – Analysis of variables collected during data collection: This analysis was not performed due to limited data on nonrespondents: gender in most cases (when the household was screened) but very rarely age for nonrespondents.

- Analysis 6 – Level-of-effort analysis: This analysis was not performed due to the lack of consistent paradata on the number of visits per case. Date of interview could not be used because the sample was released in batches (and it was more of a drip-feed approach in Northern Ireland). The alternative approach the United Kingdom took was to report the correlation between the effective "response factor" and the proficiency scores, showing that the lower the response propensity, the lower the proficiency score. This suggests a slight upwards bias may remain in the estimates, reflecting the partial, not total ability of calibration to counter nonresponse bias.

- Analysis 7 – Range of bias:

§ England: The response rate for England was 59%. The Literacy scores' first plausible value was used to compute the range of scores within the responding sample and to predict the range of estimates for nonrespondents. For the responding sample, the minimum score was 84 and the maximum score was 409, for a range of 325. Using weighting adjustment cells, and with an extreme assumption that nonrespondents would all score at the 10th percentile within each weighting cell, and at the other extreme they would all score at the 90th percentile within each weighting cell, the predicted maximum range of the mean was computed to be 47, indicating a minimal potential for bias in outcome statistics. This is a reflection of an effective nonresponse adjustment strategy carried out during weighting. That is, even though England's response rate was low (59%), the effective nonresponse adjustment weighting reduced the potential bias in the outcome statistics to a low level. However, data users need to be cautioned that the analysis is based on assumptions about the range of proficiency scores for sampled cases that have no scores (41% of the sample).

§ Northern Ireland: The response rate for Northern Ireland was 65%. The Literacy scores' first plausible value was used to compute the range of scores within the responding sample and to predict the range of estimates for nonrespondents. For the responding sample, the minimum score was 97 and the maximum score was 419, for a range of 322. Using weighting adjustment cells, and with an extreme assumption that nonrespondents would all score at the 10th percentile within each weighting cell, and at the other extreme they would all score at the 90th percentile within each weighting cell, the predicted maximum range of the mean was computed to be 37, indicating a minimal potential for bias in outcome statistics. This is a reflection of the relatively high response rate (65%) in Northern Ireland, combined with an effective nonresponse adjustment carried out during weighting.

Data collection

The United Kingdom (England)

Based on information provided on QC forms and during monthly QC conference calls, the United Kingdom (England) generally appears to have met the original requirements as described in the PIAAC Technical Standards and Guidelines (TSG), in particular Guidelines 8.1.1B and 8.1.2A on management of field staff.

The United Kingdom (England) partially met a reduced requirement on training. For the purpose of data evaluation, countries were considered to have met the standard if they provided a minimum of 15 hours of training instead of the 30 hours required by the training programme provided by the Consortium. Interviewers were provided with about 10 hours of in-person training and were offered significantly fewer training hours than recommended on key aspects (gaining cooperation and assessment administration). However, interviewers were experienced and had previously received general interviewing techniques training and at-home project-specific training.

The United Kingdom (England) did not meet a reduced requirement on validation. Standard 10.9.3 called for the validation of 10% of cases for all (100%) interviewers, selected randomly across all dispositions. For the purpose of data evaluation, countries were considered to have met the standard if they had validated at least 7% of cases for at least 96% of their interviewers, selected randomly, across all dispositions. The United Kingdom (England) reached the 7% threshold for 20% of its interviewers. Eighty percent of interviewers were validated at less than the 7% level. However, at least 10% of cases were validated overall.

The United Kingdom (Northern Ireland)

Based on information provided on QC forms and during monthly QC conference calls, the United Kingdom (Northern Ireland) generally appears to have met the original requirements as described in the PIAAC Technical Standards and Guidelines (TSG), in particular Guidelines 8.1.1B and 8.1.2A on management of field staff.

The United Kingdom (Northern Ireland) partially met a reduced requirement on training. For the purpose of data evaluation, countries were considered to have met the standard if they provided a minimum of 15 hours of training instead of the 30 hours required by the training programme provided by the Consortium. Interviewers were provided with about 10 hours of in-person training and were offered significantly fewer training hours than recommended on key aspects (gaining cooperation and assessment administration). However, interviewers were experienced and had previously received general interviewing techniques training and at-home project-specific training.

The United Kingdom (Northern Ireland) partially met a reduced requirement on validation. Standard 10.9.3 called for the validation of 10% of cases for all (100%) interviewers, selected randomly across all dispositions. For the purpose of data evaluation, countries were considered to have met the standard if they had validated at least 7% of cases for at least 96% of their interviewers, selected randomly, across all dispositions. The United Kingdom (Northern Ireland) reached the 7% threshold for 95% of its interviewers. Five percent of interviewers were validated at less than the 7% level.

Instrument data quality

Translation

To the best of the Consortium's knowledge, the United Kingdom (England and Northern Ireland) followed the PIAAC Technical Standards and Guidelines (TSG) associated with translation and verification, in particular, Standard 6.1 for new cognitive items, Standard 6.2 for background questionnaire materials, and Standard 6.3 on linking cognitive items. All adaptations were

documented and all materials went through full verification[1] prior to the Field Test and a partial verification[2] prior to the Main Survey.

· Outcome: TSG followed/Passed

Scoring

To the best of the Consortium's knowledge, the United Kingdom (England and Northern Ireland) followed the PIAAC Technical Standards and Guidelines (TSG) associated with scoring paper-and-pencil instruments, in particular, Standard 11.3.

· Coding agreement of scoring international anchor booklets

o Core items: 98.4%

o Literacy Items: 98.8%

o Numeracy Items: 96.6%

· Scoring reliability of paper-based national booklets

o Core items: 100.0%

o Literacy Items: 100.0%

o Numeracy Items: 100.0%

Assessment data

Overall, 97.4% of respondents who completed the background questionnaire (BQ) went on to take some cognitive assessment in either computer or paper format. In the United Kingdom, 83.4% of the respondents who completed the BQ took the computer-based cognitive assessment, while 14.1% took the paper-based assessment. Across all countries, 73.5% of respondents who completed the BQ took the computer-based form of the assessment and 23.9% took the paper-based form.

Some respondents who reported having computer experience refused to take the PIAAC assessment in computer-based format. Thus, these respondents took the paper-based form of the assessment. In the United Kingdom, 4.8% of respondents who reported having some computer experience refused the computer-based assessment and took the paper-based assessment. An additional 5.8% of those who reported having some computer experience failed the ICT Core and took the paper-based assessment. Overall, across all countries, 11.8% of respondents who reported computer experience refused to take the assessment on the computer and 4.7% failed the ICT Core and were therefore routed to the paper-based assessment.

The captured data for reading components showed no anomalies in terms of accuracy and missing data. Recorded time showed similar characteristics from what was seen in the Field Test in relationship to the skill of respondents.

The assignment of cognitive modules within the Virtual Machine accurately followed the intended workflow. That is to say, the administration of Literacy, Numeracy and PSTRE modules followed the

assessment design and the adaptive routing within the Literacy and Numeracy modules was accurately implemented. Analysis also showed accurate data capture for all countries.

Coding

To the best of the Consortium's knowledge, the United Kingdom followed the PIAAC Technical Standards and Guidelines (TSG) associated with coding, in particular, Standard 11.2.

· Double coding Occupation: Standard met/Passed

· Double coding Industry: Standard met/Passed

· Comparison with Labor Force Survey: Education: Standard met/Passed

· Comparison with Labor Force Survey: Occupation: Standard met/Passed

· Comparison with Labor Force Survey: Industry: Standard met/Passed

Background questionnaire data

Background data was of very high quality for the United Kingdom. If a respondent started the interview, the likelihood that she/he provided data is at above a level of 98% with practically only one exception: Income related questions are reported either in exact monetary amounts or in broad categories. In the United Kingdom, about 93.4% of respondents reported income in exact amounts (88.9% across countries) and about 1.5% reported income in broad categories (4.1% across countries). If a respondent decided to break off the interview, the interviewer was able to collect a reason for the breakoff. The data contains about 2.0% cases with breakoff codes across countries, which indicate that the reason for breakoffs were either language related issues, reading writing issues, or disabilities. In the United Kingdom, we observed 1.4% of cases with breakoffs.

Item non-response

Overall, the average proportions of non-response (omitted or not reached) for the paper-based items were 6.3% for Literacy and 4.5% for Numeracy. In the United Kingdom, these percentages were5.0% for Literacy and 3.8% for Numeracy. Overall for computer-based items, the level of non-response was 8.3% for Literacy, 5.6% for Numeracy, and 0.1% for PSTRE. For computer-based items in the United Kingdom, the percentage of non-response for Literacy was 8.8%, for Numeracy it was 6.5%, and for PSTRE it was 0.1%.