Independent Review of the 2009 PISA Results for Ireland

Report prepared for the Educational Research Centre
(at the request of the Department of Education and Skills)

Dublin: Department of Education and Skills

8 September to 15 September, 2010

Sylvie LaRoche
Fernando Cartwright
Statistics Canada

Report from a short term cost-recovery contract at the Educational Research Centre in Ireland (ERC)

Project Title: Independent Review of the 2009 PISA Results for Ireland

Client: Educational Research Centre, Dublin

Task: Review of the PISA 2009 survey for Ireland with respect to the sampling and scaling methodology

Contract Period: From 8 September to 15 September, 2010

Location: Dublin, Ireland

Authors: Sylvie LaRoche, Fernando Cartwright, Statistics Canada

1. Introduction

The Programme for International Student Assessment (PISA) is a project of the Organisation for Economic Co-operation and Development (OECD). It is designed to assess the reading, mathematics and science literacy of 15-year-olds in participating countries. The first survey took place in 2000 and has since been repeated in 2003, 2006 and 2009. The project consortium, responsible for sampling and scaling for all years since the beginning, has been led by the Australian Council for Educational Research (ACER).

In Ireland, PISA is implemented by the Educational Research Centre (ERC) on behalf of the Department of Education and Skills (DES). The ERC, on behalf of the DES, has asked the Statistical Consultation Group at Statistics Canada (SCG) to review the PISA 2009 survey for Ireland. The work was conducted at the ERC due to the OECD's embargo on the release of data pertaining to PISA 2009.

Note that ERC has already done extensive work to review the PISA survey and is completing its own report on the subject. For the SCG assessment, most of the documentation and data used were provided by the ERC or were taken from public websites. Although the ERC provided most of the data required to perform this review, the assessment was done independently by the SCG. Due to time constraints, the SCG focused its efforts on identifying potential problems or changes in the sampling or scaling methodology that could have caused a decline in estimates of achievement, mostly in the reading achievement from 2000 to 2009.

2. Objectives

The review of the PISA 2009 survey asked by the ERC comprises the following tasks:

Task 1: Assess whether the quality of organisation and administration of PISA in Ireland has met all of the PISA technical standards

Task 2: Assess and report on the extent to which the sample drawn for PISA 2009 is representative of the population of 15-year old students and the extent to which it is comparable with the samples used for PISA 2000, 2003 and 2006. This could include determining whether an additional stratification variable used in drawing the PISA 2009 sample had an impact on the comparability with previous samples and calculate and report the extent, if any to which it may have affected the scaled scores for Ireland in reading, mathematical and scientific literacy.

Task 3: Estimate and report on the contribution, if any, that demographic changes in the Irish 15-year old student population since 2000 may have made to changes in the scores achieved in the PISA 2009 assessment.

Task 4: Report on the extent to which any other sampling-related issues may have contributed to changes in PISA scores for Irish students.

Task 5: Assess the adequacy of the sampling weights used in PISA for estimating student performance and producing trends for Irish 15-year olds.

Task 6: Advise on the stability of the trend data used in the 2009 PISA survey for Ireland. This should include but not necessarily be limited to a review of (i) the PISA test design, (ii) the choice of the Rasch scaling model versus the other models, (iii) the number and content of items (and passages) selected for linking, and (iv) the impact of the decision to use international rather than national item parameters to establish trends.

Task 7: Report to the extent to which other factors relating to the PISA test design, scaling, equating, conditioning etc. may have contributed to changes in PISA scores for Irish students.

Task 8: Draw conclusions from the review and make recommendations.

The assessments made for tasks 1 to 5 are presented in Section 4 while tasks 6 and 7 are described in Section 5. A summary of findings is presented in Section 6 and recommendations are made in Section 7.

3. Overview of the PISA sampling methodology

PISA uses a systematic probability proportional to size sample design. This sample design has been used since 2000. The first-stage sampling units consist of individual schools that enrol 15-year old students. In Ireland, as in most countries, the sample of schools is drawn from a comprehensive national list of all eligible schools and the measure of size used is an estimate of the number of 15-year old enrolled in each school. Stratification, either explicit or implicit, is also used prior to sampling to improve the efficiency of the sample design and ensure adequate representation of specific groups in the sample. For more detail on the PISA sampling design, one can consult the 2000, 2003 and 2006 PISA technical reports (OECD 2002, 2005 and 2009).

The main aspects of the sample design used in Ireland remained mostly the same from 2000 to 2009. However, a few changes were introduced, mostly in 2009 and these will be discussed further in the next section.

4. Assessment of the sampling and weighting methodology

Assessing the quality of the organisation and administration of PISA 2009 in Ireland was not feasible as the SCG was not present during the administration of the PISA 2009 study. However, the PISA Consortium confirmed Ireland's adherence to the PISA technical standards to the ERC and it is officially acknowledged and documented in the PISA 2009 Main Study Data Adjudication Report (OECD Governing Board, 2010).

The SCG has also reviewed the PISA 2009 Quality Monitoring Report which provides details about the PISA Quality Monitoring (PQM) team's observations of the test implementation and the interview with the school coordinator. The information collected in this report can be used for sample adjudication purposes.

The quality monitoring was performed in seven schools. In general, the PQMs' observations indicate that there were no major issues regarding the test administration. However, it is worth noting that a number of notes point to the reluctance of students to participate, of some students leaving the test room before the end of the test or being impatient to finish and leave. This points to some student disengagement from PISA and a closer analysis of the "not reached" items and incomplete tests would be helpful to understand the extent to which this may have contributed to part of the decline in performance.

To assess whether the 2009 PISA sample is representative of the population of 15-year old students and whether it is comparable with the samples from the other PISA cycles, a review of the different stages of the sampling methodology was made first and is presented below. Due to time constraints, the SCG concentrated its efforts mainly on the 2000 and 2009 sampling methodologies.

Frame, coverage, target population and exclusions

The frames used for the school sampling are taken from the schools databases from the Department of Education and Skills (DES), which include the number of male and female 15-year olds in the schools. These databases include data from the school year prior to the school year of the PISA data collection (for example, data from the 1998-1999 school year were used for the school frame for PISA 2000). Table 4.1 shows estimates of the 15-year old enrolment, target population, and school and within-school exclusion rates. The figures for the 15-year olds and for the target population are very close, which is expected as school is mandatory until the age of 16.

One thing we notice from this table is the higher school exclusion rates in 2000 and 2003 compared to the 2006 and 2009 rates. The differences are mainly accounted for by schools for students with Special Education Needs (SEN). These schools are classified in Ireland as primary level educational institutions (ISCED 1) and not as second-level educational institutions. However, these schools include 15-year old students and were accounted for in the school-level exclusion rates for both 2000 and 2003. In 2003, the definition of the PISA

international target population was changed to include 15-year old students from Grade 7 or higher only. As the SEN schools are considered as part of the primary education sector, they were not part of the PISA target population starting in 2003 and therefore should not have been included as school exclusions in the 2003 results. Although it looks as if there was a change in the school exclusions, this had no impact on the defined target population in Ireland.

As for the within-school exclusions, no changes were observed in the definition of the exclusions or in the exclusion rates in 2000, 2003 and 2009 (Table 4.1). In 2006, the rate was slightly lower (1.67%) than for the other years but this would have only indirectly affected changes between 2000 and 2009 and was not investigated further.

Table 4.1: PISA Target Population and Exclusion Rates for Ireland

Year	15-year Old Enrolment	Target Population	School Level Exclusion (%)	Within School Exclusion%
2000	64370	63572	1.61	2.99
2003	58997	58906	1.42	2.87
2006	57648	57510	0.09	1.67
2009	Not available	55446	0.50	2.75

Source: PISA 2000, 2003 and 2006 Technical Reports

ISA 2009 Weighting Summary Report (Westat, 2010)

Sample selection

The sampling methodology used for the sample selection of all the Ireland PISA samples is mostly the same, apart from a few changes made in 2009.

In 2009, two noteworthy changes were introduced in the sampling design for Ireland. First, the International Civics and Citizenship Study (ICCS) also took place in the spring of 2009. The ICCS sample was selected first and then the PISA sample was drawn to avoid overlap with the first selected sample. The ICCS school probabilities of selection were capped at 0.5 to ensure that no schools would be selected in both samples. To control the overlap, the sample selection methodology was adapted from Keyfitz (1951). This methodology was also

used in PISA 2006 to control the overlap with the TIMSS and PIRLS samples for some countries (PISA 2006 Technical Report). Note that a potential impact could have been on the number of larger schools selected with "certainty" included in the sample as only half of them would have been allocated to the PISA sample (the other half would have been already allocated to the ICCS sample). Selecting only half of the bigger schools can have an impact if these schools have very different achievements and the two half samples selected are not similar. The samples were verified in that respect. There were no very large schools selected with certainty in 2000, 2003 or 2006. The sample selection with the overlap control was verified and no problems were identified in the process.

The second change was the modification of stratification variables used for the sample selection. Stratification is used in part to ensure adequate representation of specific groups of the target population in the sample and to improve the precision of the sample design. The stronger the correlation between the achievement and the stratification variables, the better the precision of the estimates is. The stratification variables used in 2000 and 2009 are described in Table 4.2. Both explicit and implicit stratification will ensure a proportional allocation across the groupings used (that is, if the allocation to the explicit strata is proportional, which is the case in Ireland).

Table 4.2: Stratification variables used in PISA 2000 and 2009 for Ireland

Stratification	2000	2009
Explicit	Size of school (17-40, 41-80, 81+)	Size of school (1-40, 41-80, 81+) Type of school (Secondary, Vocational, community/comprehensive)
Implicit	Type of school (Secondary, Vocational, community/comprehensive) School gender (Girls only, mixed, boys only)	Percentage of students in school with a Junior Certificate fee waiver (4 categories based on quartiles) Percentage of female in school (4 categories: 0%, more than 0% but less that 45%, 45% to less than 100%, and 100%)

Source: Sampling Forms for PISA 2000 and 2009 sampling and internal documentation

In 2009, the percentage of students in school with a Junior Certificate fee waiver[1], which is an indication of the socio-economic status in the school was introduced as a first implicit stratification variable for the sample selection (a low percentage of Junior Certificate fee waiver indicates a higher socio-economic status, i.e., a more advantaged school). Use of this variable, developed in 2005, was designed to ensure a proportional allocation of schools across the four quartiles of the "% fee waiver" in 2009 (assuming this percentage is a fairly stable for each school from 2005 to 2009). Since this variable was not used in 2000, an attempt was made to verify the distribution of this variable in the 2000 sample. One limitation to this verification is that the distribution of the population estimates for this variable is not available for 2000. However, since the data used to derive this stratification variable in 2009 were based on information averaged over 2002 to 2004 (and is thought to be fairly stable, according to the ERC), the distribution used for 2009 was also considered valid for comparison with the 2000 sample distribution. The verification showed that there were no significant differences between the estimates from the 2000 sample and the ones from the "2000 population estimates" for this variable.

Because the frame is sorted by implicit strata prior to selection, using the "% fee waiver" in stratification could have affected the choice of replacement schools used for non-participating schools. In 2009, the only 2 replacement schools used were from the same "% fee waiver" quartile as the original schools they replaced. Since the "% fee waiver" was not used in the 2000 stratification, the replacement schools used may have been from a different "% fee waiver" quartile. Only three replacement schools were used in 2000 and results show there were no significant differences between the "% fee waiver" of the original schools and their replacements. Therefore, using the "% fee waiver" for implicit stratification does not appear to be a cause for concern.

Comparisons made between the 2000 and 2009 sample estimates with their respective population estimates with respect to all the stratification variables show that the samples are comparable to the corresponding populations from which they were drawn and are valid. All

[1] The fee waiver percentage is taken from the 2005 data which is derived from an average over 3 years (2002 to 2004). Note that the lower the percentage of Junior Certificate fee waiver is, the more advantaged the school is.

of the 95% confidence interval limits cover the corresponding population, as expected under unbiased sampling and random non-response. Extensive comparisons were also made by the ERC and are described in their report (Cosgrove et al., 2010).

The verification performed on the sample selection for 2000 and 2009 show that processes used by the PISA Consortium were valid and did not cause problems.

Demographic changes

Demographic changes were observed in the estimates from the 2000 and 2009 samples. Nothing was found in the sampling and weighting methodology that could have created these changes. Therefore, these observed changes are assumed to be a reflection of true changes in the population composition between 2000 and 2009. The ERC has already analysed and documented these changes and due to a lack of time, the SCG was unable to look at these demographic changes further.

Weighting procedures

The derivation of the survey weights in PISA is done according to the standards for the complex survey data analysis (PISA 2006 Technical Report). Generally, the same procedures for deriving the sampling weights are used in all PISA cycles and in other international studies. Between 2003 and 2006, changes in the student non-response adjustment procedure were introduced to better account for difference in non-participation patterns by grade and gender. These changes are documented in the PISA 2006 Technical Report and an evaluation made by the PISA Consortium showed the impact of that change was of less than 2 points in the achievement score in all countries. Due to time constraints and complete weighting files not being readily accessible, it was not possible to examine this matter further.

The sampling weights and weight adjustments were examined to detect any unusual or important fluctuation that could have had an impact on the estimated mean achievement scores in 2000 and 2009. The distributions of the school and student base weights and of all

the weight adjustments (school grade adjustment, non-response adjustment for school and students) were verified and there was no evidence pointing to problems in the weighting of the 2000 and 2009 samples.

Sample weights are adjusted to account for the non-response of schools and students. At the school level, the stratification variables are used to form groups of schools with similar response patterns. Another aspect that was examined was the inclusion of the "%fee waiver" variable as a stratification variable in 2009 and therefore, its use in the school non-response adjustments strategy. Since the "% fee waiver" is a good predictor of non-participation and is well correlated to the school achievement, using this variable to form non-response groups can only reduce a potential non-response bias and improve the precision of the estimates.

Since that variable was not used in 2000 to adjust for the school non-response, a review of the 2000 non-participating schools was made to verify the impact, if any, of not using the "% fee waiver" variable in the 2000 non-response adjustment. The results showed that 50% of the non-participating schools in 2009 were from the bottom "% fee waiver"[2] quartile and another 39% were from the next lowest quartile. Since the schools with lower "% fee waiver" (more advantaged schools) are likely to be better achieving schools, incorporating the "% fee waiver" into the non-response adjustment in 2000 may have had the impact of increasing the achievement scores (assuming that the variables used for the school non-response adjustment in 2000 were not as well correlated with non-response and achievement as the "% fee waiver").

"Outlier achievement schools"

Although a more global decrease in achievement is observed in the 2009 data, seven schools[3] were identified as having a reading achievement score of more than 100 points (one standard deviation) below the mean score in 2009. No such schools were found in 2000. Three of these schools particularly stand out as extreme cases. A careful analysis was made of the

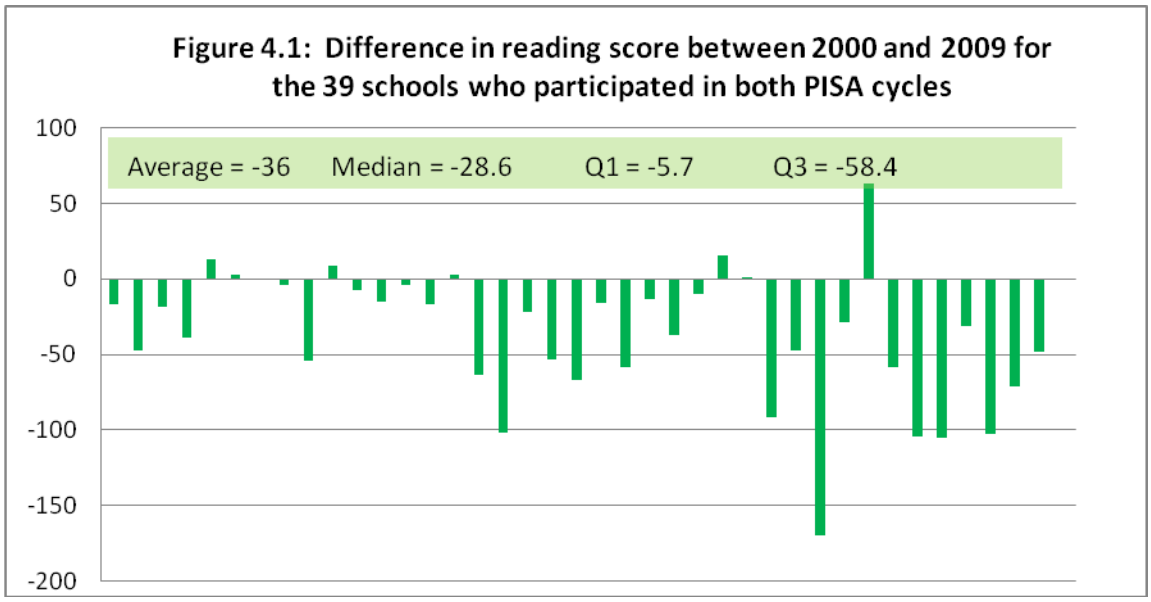
[2] The lower the percentage of Junior Certificate fee waiver is, the more advantaged the school is.

[3] The ERC's report indicates that there were 8 such schools in 2009 but their estimates are based on unweighted data using the first plausible value.

school and student weights as well as all the weight adjustments for these low achieving schools to see if there were extreme adjustments made to the weights. No problems with the weights or adjustments were found. Student participation for all these schools was in line with what was found in other schools. More analysis is required to identify the reasons why the scores of these schools are much lower and the ERC is looking into this in more depth.

Comparison of the schools found in both the 2000 and 2009 samples

Thirty-nine schools were identified as being in both the 2000 and 2009 samples. Reading scores from 2000 and 2009 were compared for the schools (Figure 4.1). These schools show a decline of 36 point on average. However, the median show a decrease of 28.6 points. The difference between the average and the median is caused by the extreme drop in the reading score of one particular school (-170 points). Therefore it is better to use the median to compare the decline of these schools with the national decline and the decline by 28.6 points is similar to the decline observed nationally (-31 points). Again, for these schools, all weights and adjustments were carefully examined and nothing was found in the sampling methodology that could explain the decline.



Again, due to time constraints, it was not possible for the SCG to analyse further the data for those schools that participated in both 2000 and 2009 but it would be also interesting to see how the achievement scores in mathematics and science from 2000 and 2009 compare. Another interesting analysis would be to also identify schools that participated in the 2003 and 2009 cycles and in the 2006 and 2009 cycles and to do similar comparison of the scores.

Participation Rates

Participation rates from 2000 to 2009 were compared to see if there were significant changes in participation, either at the school or student levels. In general, participation rates remain fairly similar apart from the 100% school participation rate in 2006 (see Table 4.3). Both the school participation and student participation are above the OECD requirements for minimum participation and meet the sampling standards. However, the student participation rate in Ireland has been systematically below the average over all countries since 2000 (by 4 percentage points in 2000 and close to 9 percentage points in 2003 and 2006). It is important to keep in mind that higher non-participation increases the risk of non-response bias.

Table 4.3: Participation Rates in PISA 2000, 2003, 2006 and 2009

	Number of participants	School Participation Rate (weighted)	Student Participation Rate (weighted)	Average Student Participation Rate (weighted) - All countries
2000	3854	87.5%	85.6%	89.8%
2003	3880	92.8%	82.6%	91.4%
2006	4585	100%	83.7%	92.5%
2009	3896	88.4%	83.8%	Not available

Source: PISA 2000, 2003 and 2006 Technical Reports

PISA 2009 Weighting Summary Report (Westat, 2010)

At the school level, more variables that are correlated with achievement, are available to perform non-response adjustment (school size, school type, the "% fee waiver" and the proportion of girls in school). At the student level, variables that can be used for non-response adjustment are limited to gender and grade. Although these variables explain some of the non-response and are correlated with the achievement, there might be other unknown characteristics correlated with achievement that distinguish the non-participating students

from the ones who participate. Increasing student participation to PISA would reduce the risk of non-response bias and would be beneficial. Make-up sessions for students who did not participate in the test could be used to increase participation rates.

5. Assessment of the scaling

PISA test design

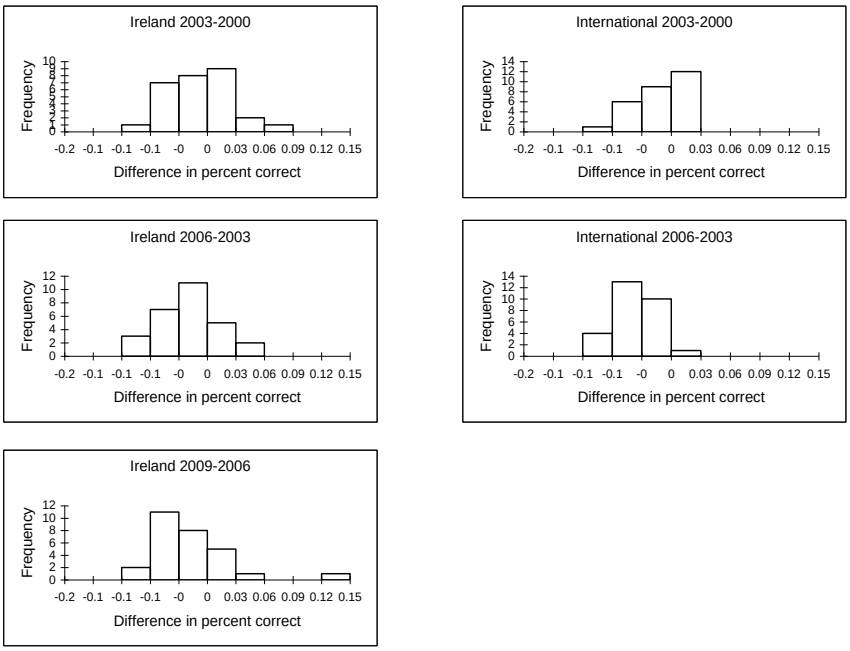
The PISA test design consists of blocks of items, which are rotated through test booklets. Each booklet consists of four blocks. In PISA 2000, the reading items were placed solely in the first three blocks. In subsequent PISA cycles, the reading items were balanced among all four blocks (though often preceded and followed by different domains/subjects). The effects of item position in PISA 2000 are documented in the PISA Technical Report (OECD, 2002). The reading domain moved from a major domain in PISA 2000 to a minor domain in PISA 2003 and PISA 2006, finally returning to a major domain in PISA 2009. In addition to these structural changes, several reading units were changed from their initial design between 2000 and 2003 (e.g., items were omitted). Since these changes were implemented universally, it is not possible objectively to determine the consequences of these item adjustments or adequately model them statistically. However, some consistent patterns of item response can be observed in the subsequent cycles (Table 5.1, Figure 5.1). Figures are not available for the international sample, because the item data are not yet released for all countries. These values differ from the values produced in the national item reports, because these figures are based only on the common items and countries.

Table 5.1: Proportion correct, common items, Ireland and international averages, PISA 2000, 2003, 2006 and 2009

Item Name	IRL_2000	IRL_2003	IRL_2006	IRL_2009	INT_2000	INT_2003	INT_2006
R055Q01	0.88	0.88	0.87	0.83	0.86	0.83	0.82
R055Q02	0.57	0.59	0.58	0.56	0.53	0.48	0.46
R055Q03	0.73	0.69	0.66	0.66	0.62	0.59	0.58
R055Q05	0.84	0.82	0.78	0.75	0.77	0.72	0.68
R067Q01	0.93	0.93	0.92	0.91	0.89	0.90	0.88
R067Q04	0.75	0.77	0.81	0.79	0.72	0.74	0.73
R067Q05	0.78	0.84	0.83	0.79	0.71	0.74	0.73
R102Q04A	0.38	0.30	0.32	0.28	0.38	0.33	0.27
R102Q05	0.51	0.52	0.54	0.49	0.41	0.44	0.41
R102Q07	0.92	0.90	0.91	0.92	0.87	0.84	0.84
R104Q01	0.88	0.90	0.87	0.81	0.83	0.83	0.78
R104Q02	0.43	0.38	0.34	0.38	0.42	0.35	0.32
R104Q05	0.52	0.49	0.42	0.33	0.49	0.44	0.38
R111Q01	0.69	0.69	0.66	0.63	0.66	0.67	0.63
R111Q02B	0.67	0.70	0.64	0.62	0.51	0.52	0.53
R111Q06B	0.52	0.58	0.52	0.51	0.52	0.51	0.48
R219Q01E	0.68	0.68	0.63	0.60	0.56	0.58	0.52
R219Q01T	0.87	0.87	0.83	0.83	0.69	0.70	0.64
R219Q02	0.90	0.88	0.89	0.85	0.75	0.77	0.75
R220Q01	0.48	0.44	0.43	0.41	0.47	0.45	0.39
R220Q02B	0.62	0.63	0.63	0.62	0.65	0.65	0.60
R220Q04	0.56	0.61	0.55	0.52	0.62	0.62	0.58
R220Q05	0.88	0.86	0.83	0.78	0.86	0.84	0.79
R220Q06	0.62	0.60	0.61	0.57	0.68	0.68	0.65
R227Q01	0.46	0.41	0.46	0.44	0.61	0.56	0.48
R227Q02T	0.88	0.85	0.80	0.93	0.83	0.81	0.73
R227Q03	0.59	0.55	0.53	0.54	0.56	0.55	0.52
R227Q06	0.81	0.77	0.72	0.75	0.75	0.73	0.69

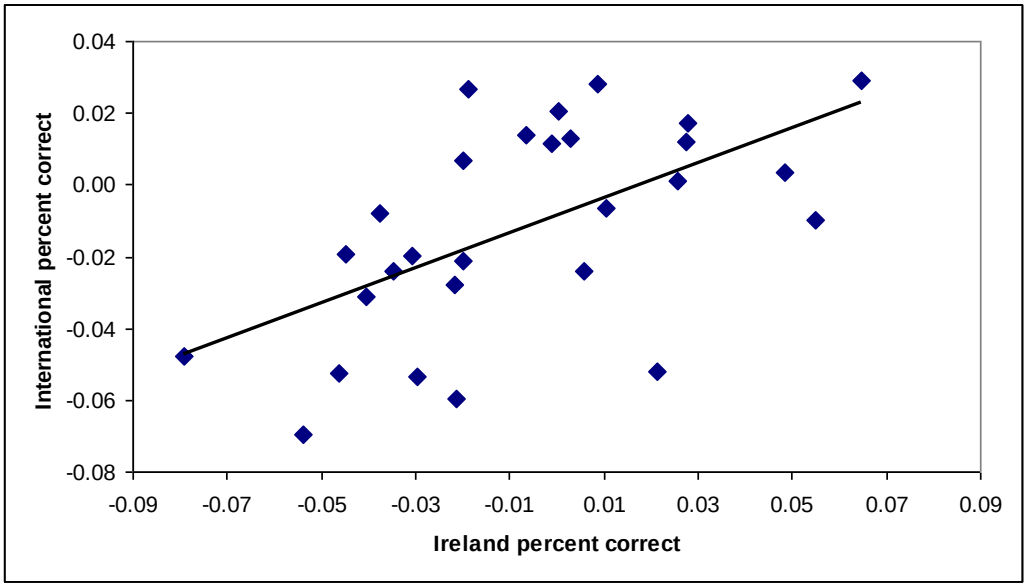
The international average was computed on the basis of countries common to all PISA cycles.

Figure 5.1: Distribution in changes in item proportion correct for common items Ireland and international average, PISA 2000, 2003, 2006 and 2009



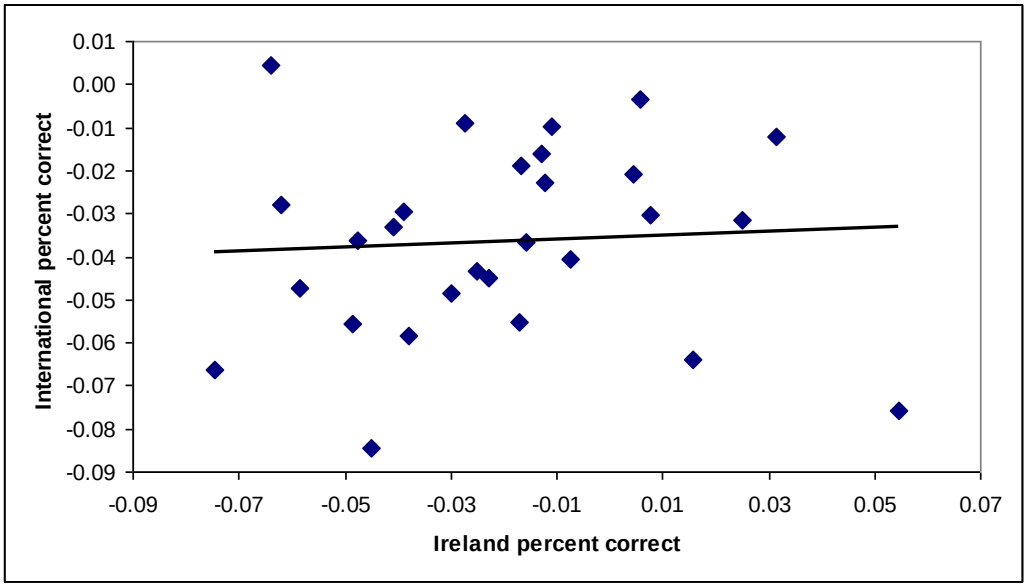
In each subsequent cycle, a smaller proportion of students has been answering the link items correctly in both Ireland and the international sample (excluding Ireland). Where there is a difference between the decrease in Ireland and the international sample, the international sample had a larger decrease. These results suggest that, while there may be a decline in student performance in Ireland, it is consistent with a decline in performance across all PISA participants. For example, the relationship between the change in performance for Ireland between 2000 and 2003 is compared with the changes in the other participating countries in Figure 5.3. The generally positive trend indicates that the average change is not dominated by a single item; rather, items that tend to have a lower probability of being correctly answered in Ireland also tend to have a lower chance of correct response in other countries. This consistency suggests that the changes may be structural in nature, rather than the result of changes in underlying student proficiency in individual countries.

Figure 5.2: Relationship between changes in percent correct for Ireland and other PISA participants, PISA 2003-PISA2000.



In contrast, the changes in item performance between 2003 and 2006 do not have similarly consistent patterns (Figure 5.3). These results are consistent with the structural explanation for the changes between PISA 2000 and 2003. There were no structural changes to the reading assessment between PISA 2003 and PISA 2006. Therefore, differences in item performance between cycles are more likely to be sensitive to random and country-specific factors.

Figure 5.3: Relationship between changes in percent correct for Ireland and other PISA participants, PISA 2003-PISA2000.



One issue affecting comparisons of performance over time is the random equivalence of the link items to the non-link items. Although is it mathematically possible to construct a linkage using a single item, such a linkage would be very unstable. The validity of the linkage depends on representation of items for each level of proficiency for which the linkage will be generalized. A summary comparison of the per cent for the link items and non-link items over time, for Ireland and other common countries (Table 5.2) illustrates the imbalance of linking items towards easier items. This imbalance is noted in the PISA 2003 Technical Report. However, with no control sample, it is difficult to draw conclusions on the effect of this imbalance on the quality of the linkage. There is no evidence to suggest that the linkage is less stable at the higher end – indeed, it would appear that the link is less stable at the lower end, for the items that do have a linkage. Of the 26 link items between PISA 2006 and PISA 2009, 11 items were more difficult than average in 2009 and 15 were easier than average. Among the 11 difficult items, the correlation between the national item difficulties in Ireland and the international item difficulties (produced for the national item reports) is 0.79, whereas the corresponding correlation for the easy items is 0.92.

Table 5.2 Item percent correct for link and non-link items, Ireland and other common participants.

Ireland				
	2000	2003	2006	2009
Non-link	0.67	…	…	0.60
Link	0.69	0.68	0.66	0.65
Other participants				
	2000	2003	2006	2009
Non-link	0.63	…	…	…
Link	0.65	0.64	0.60	…

Note. These statistics differ from the results published by the PISA Consortium as a result of 1) treating not-reached items as missing, and 2) restricting the international population to those countries with continuous participation in PISA since 2000 (excluding Ireland). Individual item response data are not available for PISA 2009 for the international sample.

Although there may be issues with the PISA test design that produce biased or error-laden estimates of the differences in performance between PISA cycles, these errors would be consistent across all participating PISA countries. For example, if the consistent gap in PISA performance seen among high performing countries between PISA 2000 and PISA 2003 is artefactual (although this hypothesis cannot be tested with existing data beyond mentioning the unlikelihood of a internationally consistent patterns of decreasing item performance), this error would be applied consistently to all countries and cannot account for the relative drop in performance of Ireland compared to other countries. However, this potential artefact should be considered if performance is to be interpreted against static benchmarks, such as the PISA 2000 international average, which have not been affected by any possible structural changes.

Choice of the Rasch scaling model

The Rasch scaling model makes strong assumptions about equivalence across items and populations. This model is the most restrictive of the available item response scaling models, and is therefore the most prone to misfits between the model and the data to which it is applied. Model data misfits should be random within the population used to calibrate the models, since the estimation procedures tend to minimize average error. However, for specific subpopulations or specific items, the errors may be quite large, which may

unfairly bias estimates. To analyse the fit of the model to the data, we compared the empirical estimates of probability of students scoring an item correct as a function of their proficiency (estimated using the first plausible value) to the predicted probability based on the international item parameters for the PISA 2009 response data in Ireland. Students were split into 25 groups defined by equally spaced 25-point intervals on the PISA scale in order to estimate group probabilities. The difference in probability was calculated separately for each group using sample weights and then averaged, unweighted, across groups. The resulting statistic (Table 5.3) describes the average model data misfit, rather than a summary of population difference in performance from the expected values. For ease of communication (i.e., single summary statistic), this analysis was conducted using dichotomous items only (i.e. excluding partial credit items).

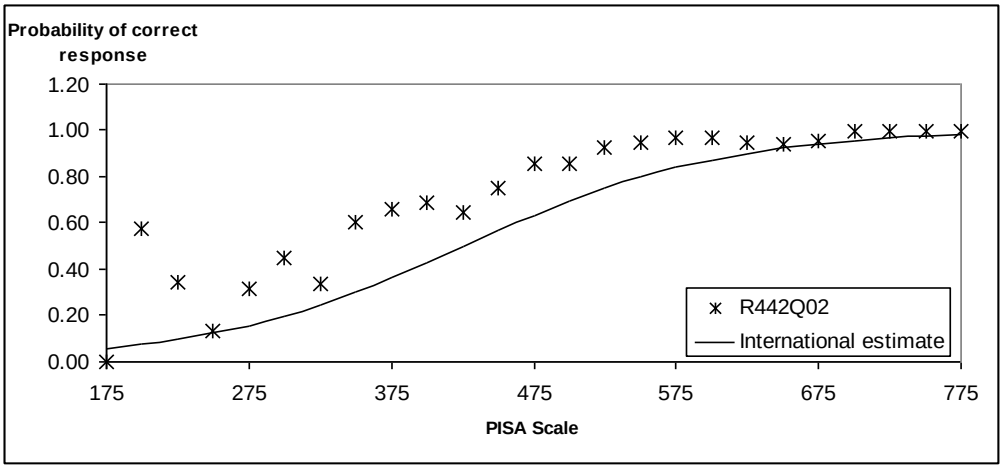
Table 5.3: Differences between expected and observed probability of correct response for dichotomous items, PISA 2009, Ireland.

Name	International difficulty	On PISA Scale	Average Δ Probability	Name	International difficulty	On PISA Scale	Average ΔProbability
R055Q01	-1.459	349	0.012	R420Q02	-1.583	339	-0.062
R055Q02	0.474	504	0.008	R420Q06	0.768	528	-0.011
R055Q03	0.012	467	0.030	R420Q09	-1.037	383	0.012
R055Q05	-0.771	404	-0.032	R424Q02T	0.958	543	0.071
R067Q01	-2.135	295	0.030	R424Q03	-0.325	440	0.028
R083Q01	0.072	472	0.041	R424Q07	-0.912	393	0.057
R083Q02	-1.382	356	0.035	R432Q01	-1.572	340	0.063
R083Q03	-1.049	382	-0.014	R432Q05	-0.660	413	0.049
R083Q04	-0.307	442	0.030	R432Q06T	2.890	697	-0.119
R101Q01	0.486	505	-0.033	R437Q01	0.335	493	-0.011
R101Q02	-1.457	350	0.019	R437Q06	0.333	493	0.078
R101Q03	0.028	468	0.029	R437Q07	2.483	665	-0.049
R101Q04	-1.025	384	0.058	R442Q02	-0.501	426	0.140
R101Q05	0.808	531	0.032	R442Q03	-0.592	419	0.050
R102Q04A	1.422	580	-0.037	R442Q05	1.345	574	-0.079
R102Q05	0.668	520	0.011	R442Q06	1.969	624	0.006
R102Q07	-1.547	342	0.071	R442Q07	1.196	562	-0.035
R104Q01	-1.255	366	0.041	R446Q03	-2.484	267	0.040
R104Q02	1.330	572	0.058	R446Q06	-1.045	383	0.084
R111Q01	-0.377	436	-0.023	R447Q01T	-0.251	446	0.095
R219Q02	-1.423	352	0.023	R447Q04	0.330	493	-0.028
R220Q01	0.901	538	-0.034	R447Q05	-1.037	383	0.045
R220Q02B	-0.179	452	0.019	R447Q06	0.687	521	0.004
R220Q04	-0.005	466	-0.027	R452Q03	2.918	700	-0.040
R220Q05	-1.347	358	-0.029	R452Q04	-0.266	445	-0.007
R220Q06	-0.435	431	-0.082	R452Q06	0.633	517	0.037
R227Q01	0.172	480	-0.072	R452Q07	0.760	527	0.028
R227Q03	0.225	484	-0.028	R453Q01	-1.307	362	0.066
R227Q06	-0.888	395	-0.010	R453Q04	-0.159	453	0.113
R245Q01	-0.377	436	-0.009	R453Q05T	-0.170	453	0.042
R245Q02	-0.415	433	0.025	R453Q06	-0.611	417	0.048
R404Q03	-0.757	406	0.059	R455Q02	1.245	566	0.076
R404Q06	0.654	518	0.020	R455Q03	-1.041	383	0.035
R404Q07T	1.436	581	-0.058	R455Q04	-0.251	446	0.001
R404Q10A	0.940	541	0.019	R455Q05T	1.960	623	0.052
R404Q10B	1.189	561	-0.003	R456Q01	-3.396	194	-0.005
R406Q01	-0.339	439	0.020	R456Q02	-1.380	356	0.111
R406Q02	1.426	580	0.040	R456Q06	-1.365	357	0.059
R406Q05	-0.763	405	0.081	R458Q01	0.519	508	0.009
R412Q01	-1.650	334	0.071	R458Q04	0.095	474	0.001
R412Q05	0.111	475	-0.029	R458Q07	0.237	485	-0.061
R412Q06T	1.114	555	0.042	R460Q01	-0.295	443	0.059
R412Q08	1.052	550	0.043	R460Q05	-1.369	357	0.078
R414Q02	0.817	531	0.052	R460Q06	0.003	466	-0.012
R414Q06	0.379	496	0.036	R466Q02	0.855	535	-0.022
R414Q09	-0.092	459	0.078	R466Q03T	2.662	679	-0.062
R414Q11	1.355	575	-0.018	R466Q06	-1.171	372	-0.040

Note. Difference is calculated as Ireland observed probability minus expected probability based on international parameters. At the time of this study, the published item parameters on the PISA scale were not available, so the conversion constants were taken to be consistent with the reported linking methodology, which converts PISA 2009 to the PISA 2006 scale then adopts the PISA 2006 link to PISA 2000. We used the not-reported gender constants: ((0.8830 * 0.0906 * Logit + 0.0552 – 0.5076) / 1.1002) * 100 + 500.

The results of this analysis indicate a fairly consistent bias in Ireland with respect to model data fit. Performance in Ireland is better than would be expected on 65% of the items, given the PISA scores assigned to the students. The largest differences occur in the items unique to PISA 2009. For the most part, the linking items fit the model reasonably well. For illustration, Figure 5.3 shows the results for item R442Q02, which had the greatest average difference (0.140). In this figure, the asterixes represent the empirical estimates based on observed responses, and the curve represents the expected performance based on the international parameters. Across almost every proficiency group, students in Ireland performed better than their estimated scores describe.

Figure 5.3: Differences between expected and observed probability of correct response for item R442Q02, PISA 2009, Ireland.



The consequences of this bias in the model data fit are that the PISA scores are not adequately representing the performance of students in Ireland. For example, in Figure 5.3, the literal interpretation of the PISA scale scores using the international item parameters suggests that students in Ireland with scores around 375 will answer this item correct

approximately 30% of the time. However, observed data indicate this value is closer to 60%. Inferences about actual student performance based on the PISA scores will tend to underestimate what and how well students can actually perform.

The reasons for these model-data fit issues may come from two sources. The first is inappropriateness of the statistical model. That is, assuming items are more accurate than they truly are tends to produce bias in model data fit for certain subpopulations, depending on if their performance is above or below the average. Generally, performance will be underrepresented in subpopulations with above-average item performance, such as Ireland when the accuracy of the items are overestimated by the model. This limitation can typically be circumvented by allowing the statistical item model to describe the conditional accuracy of the item (with respect to proficiency) as variable across items. This approach is used in many other international studies and has been recommended for PISA as well (see Mazzeo & von Davier, 2008). However, the costs of adopting this approach may include losing the capacity to measure trends at all with the existing published data.

Another possible cause of the bias in model-data fit is item by country interactions. It is not possible to examine the impact of country interactions in the absence of the international response data files. This issue is discussed in a later section of this report.

Number and content of items selected for linking

As mentioned earlier, the validity of a linkage depends in large part on the representativeness of the linking anchor test of the domain being measured. For PISA reading, given the broad and novel approach taken to the definition of reading by PISA, there is no natural 'curriculum' from which to define the domain. However, we can use the initial design and composition of PISA 2000 as an adequate representation of this ideal domain definition. Table 5.4 summarizes the composition of the PISA 2000 original assessment and the anchor test used for linking, which is a subset of the original test. The items are classified according to their reading sub-domain and an item difficulty grouping (based on the item percent correct in PISA 2000). Each cell in the table describes the

representation of the items in each test as a percentage. The comparison suggests that, at least in terms of these characteristics, the proportional composition of the anchor test accurately reflects the composition of the original PISA 2000 test (notwithstanding, as discussed earlier, the overall easier content in the anchor test).

Table 5.3: Composition of original PISA 2000 reading and anchor tests (percentages of items)

	Item Difficulty Group				
Content	1	2	3	4	Total
PISA 2000					
InterpretingTexts	1	9	25	15	49
Refl.&eval.	2	8	9	4	23
Retr.inf.	1	5	9	13	28
Total	4	22	43	32	100
Anchor test (2003-2006)					
InterpretingTexts	0	7	25	18	50
Refl.&eval.	0	7	14	4	25
Retr.inf.	4	7	7	7	25
Total	4	21	46	29	100

However, in terms of the number of items and passages used for the linkage, the anchor test does not reflect the accuracy of the original. It is unclear whether the information available from these items is sufficient to establish a strong linkage. For example, the PISA 2006 Technical Report (OECD, 2009) provides an estimate of the reading anchor test reliability of 0.429. Although the final estimates (after multidimensional scaling and conditioning) have a much higher reliability, it is extremely important to remember that the final scaling with these additional sources of data is performed nationally. Therefore, the national averages do not benefit from any increased precision that results from the inclusion of additional information; the national averages are still produced using a test with a reliability of 0.429. This level of reliability is generally considered too low for any practical purpose. For general research purposes, this level of reliability would reduce the observed strength of correlation between the test scores and other variables to 65% of their 'true' value. On a simple test-retest scenario using a test with this level of reliability, if a student scores 0.5 standard deviations above the average, the same student would expect to score 0.21 standard deviations above the average on the retest, *even if their underlying*

proficiency had not changed. This phenomenon, known as regression to the mean, tends to exaggerate drops in performance for initially high performers as well as increases in performance for low performers due to the nature of random errors. It is not possible to quantify precisely the effects of this phenomenon with the PISA data in Ireland in the current study due to the changes in test accuracy between PISA cycles and the ambiguity between countries versus students as the unit of measurement.

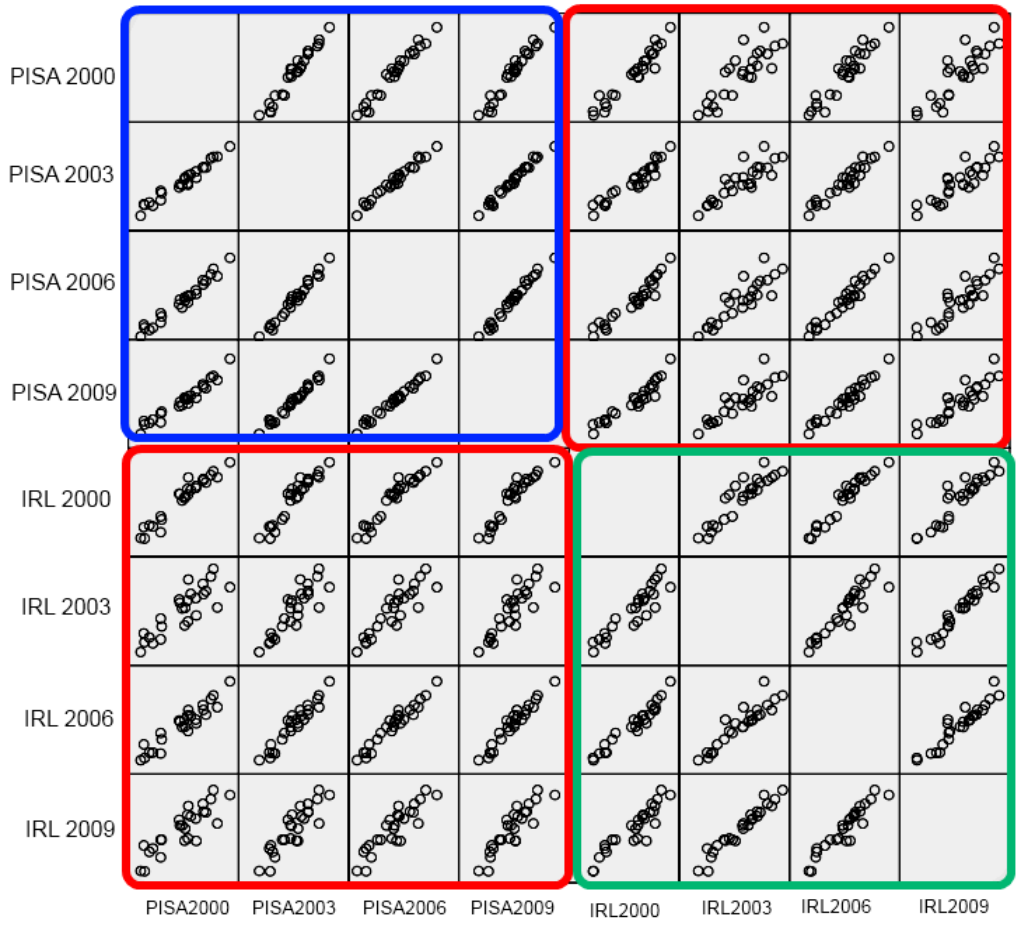
Although it appears performance is consistently decreasing in Ireland, regression to the mean always affects interpretation of test-retest results when the instrument is less than perfect. This phenomenon is unavoidable, but it is also random, which means we cannot disentangle for a single country how much of an observed change is due to regression to the mean. We simply know that, on average, initially high countries tend to decrease and initially low countries tend to increase, and the effect increases as the inaccuracy of the instrument increases. However, the inaccuracy of the linking anchor test suggests that any drop in performance in PISA by initially high performing countries (such as Ireland) or increase in performance of initially low performing countries should be considered with caution and corroborated by independent data sources.

Using international rather than national item parameters to establish trends

The issue of using national versus international item parameters to establish trends has been discussed in the PISA 2003 Technical Report (OECD 2005) and examined in detail by Gebhardt & Adams (2007). The latter study performed a re-analysis of PISA data specifically to examine the consequences of establishing the linkages at a country level, rather than internationally. The main results are that the country-estimated trends deviate substantially from the trends estimated using the fixed international linkage. The trend for Ireland using the national linkage was relatively similar to the trend estimated using the international linkage in their study for PISA 2000 to PISA 2003. However, due to the random nature of these variations, it is likely that deviations in subsequent cycles may be more reflective of the general patterns of variation seen in other countries.

The challenge in estimating national trends arises from the errors in estimating item parameters. The PISA major-minor design is a challenge to this issue, because the sample size for link items in minor domains is relatively small, and the estimates rely on very sparse measurement information (recall that each student is administered *fewer* than 28 items). As a result, the parameter estimates for items are more unstable year-to-year within a country than they are at the international level. The scatterplot matrix in Figure 5.4 illustrates the stability of the item parameter estimates over time for Ireland and the PISA international calibration sample for the linking items. The blue submatrix in the upper left shows the year-to-year comparisons of the international calibration samples. These estimates are very stable, with almost perfect correlations between adjacent cycles. In contrast, the red submatrics in the upper right and bottom left (they are mirror images or each other) display the relationships between the international and national item parameters. There is a great deal of imprecision, even within the same PISA cycle. However, it is difficult to conclude that the weaker relationship between international and national estimates is due to systematic differences between Ireland and the international sample or simply due to random error. After all, the estimates within Ireland (bottom right submatrix) have as much variation between adjacent cycles as the Ireland-to-international comparisons in the same cycle. Without an objective source of 'truth' the following possibilities are potentially valid: 1) the international parameters are correct and the lack of fit in Ireland is due to sampling error, 2) the international parameters are biased and do not represent the true relationships in Ireland, and 3) some combination of the two.

Figure 5.4: Relationship between Irish and international item parameter estimates over time, PISA 2000 to PISA 2009



However, discussion of estimating linkages at the national level raises the question about the purpose of PISA and how to use the data most effectively. Estimating trends at a national level may provide a more accurate picture of change, but it weakens the comparability between countries. In other words, the increase in accuracy within Ireland would mean that comparisons of changes between Ireland and other countries are less accurate. Given the role of PISA specifically as an international assessment intended to facilitate international comparisons, the benefit of effectively converting it into a tool primarily for estimating within-country results is dubious, particularly given the rich national assessment systems already in place in Ireland. Given this consideration, and the fact that the international parameters are generally within the sampling confidence intervals of the national estimates, the most useful course of action is to retain the

international parameters. The increased stability of these estimates comes at the cost of potential bias, but this cost is less than the instability of the national estimates, which would introduce too much random error to make the trends interpretable.

Other factors contributing to changes in PISA scores for Irish students.

There are two main factors that have potentially the largest impact on the interpretation of the current PISA 2009 results for Ireland. The first is the conceptual gap between student proficiency and student performance, and the second is the imprecision or inadequate representation of linking error in PISA.

Student engagement

Student performance refers to the observed behaviour of students on test items. Student proficiency, on the other hand, is the trait that students use to generate that behaviour. The scores produced for a test, such as the PISA plausible values, use student performance to describe student proficiency. However, although standardized test design and administration are intended to bridge the gap between the two, it is important to remember that they are not the same. Specific to the current PISA results, there are two issues in particular that appear to be increasing the difference between the two. The first issue is student engagement with PISA and the second is the PISA scoring methodology.

Over time, the percent of items that students have not attempted (either due to missing or not reached) has increased consistently since 2003 (Table 5.4). These changes in response patterns are primarily due to increases for lower performing students, but the phenomenon has happened across most student populations in Ireland. Although it is possible that increasing item level non-response may be the result of decreasing proficiency, in the absence of corroborating evidence to confirm such a decrease, the consistent pattern of missing responses, not-reached items, and poorer performance on linking items (see Table 5.2) suggests a more pervasive pattern of declining performance that may be specific to the PISA assessment. Missing responses and not-reached items in particular are evidence of respondent fatigue or lack of engagement with the assessment process than they are of poor proficiency. Unfortunately, it is beyond the scope of this study to examine the school

level factors and demographics that may differentiate those schools or individuals who are more likely to be associated with item-level non-response.

Table 5.4: Percent of Missing or Not Reached item responses across all items for Reading Literacy, Ireland, PISA 2000 to 2009

	PISA 2000	PISA 2003	PISA 2006	PISA 2009
Missing responses	4.9	5.7	6.2	7.9
Not reached items	1.9	0.6	0.8	2.4

Regardless of the reasons why individuals are not providing item responses, the consequences are magnified during the scoring process, because not-reached and missing responses result in a score of 0 during the PISA scoring. In other words, despite an absence of information regarding whether a student can or cannot perform a task, the process infers from the missing response that the student cannot perform the task. Invariably, the result of this scoring approach is to increase the bias in final estimated scores downwards. The degree of bias increases with the proportion of missing responses. Alternate approaches, such as treating item-level non-response as missing may be equally problematic, since students may artificially inflate their scores by only answering items they know with certainty. The best compromise is to treat missing items (i.e. those preceded and followed by items the student *has* attempted) as incorrect, while leaving unreached items as missing. The scores produced by this approach tend to have larger measurement errors, because they are based on fewer item responses, but they are also less prone to bias. As fatigue with PISA administration increases over time with schools and teachers, it may become necessary to alter the scoring methods to avoid permanently confounding performance issues that are related to administration conditions with changes in student proficiency.

Linking error

Possibly the most telling statement regarding the representation of linking error in PISA is made by Gebhardt & Adams (2007): "…no consensus has been reached about [estimation of linking errors]" (p. 309). Currently, the process of estimating linkages for PISA has

been introduced and changed twice. First, the method used for PISA 2000 to PISA 2003 was found to underestimate the link error (Monseur & Berezner, 2007). The authors of that study proposed an alternate method that accounted for clustering of items, resulting in larger estimates of linking error. In addition, the authors proposed additional changes to the methodology, including accounting for the larger contributions of partial credit items, reporting country-specific linking errors, and describing the effects of model misspecification on the linkages.

The most recent change to the methodology incorporated in PISA has been the introduction of what is described on page 159 of the PISA 2006 Technical Report (OECD, 2009). The description of method in this text appears to be incomplete, as it seems inconsistent with the reported values. The description of method also appears to have several inconsistencies or typographical errors in notation and provides no clear description of how a single linking error can be estimated, despite having scale inconsistencies for males and females. However, it is unlikely at this time that these issues will be clarified.

In order to provide a basis for discussion for the current analysis, the linking errors reported by PISA are compared in Table 5.5 to estimates that are consistent with the methodological priorities described in the Technical Report. The derivation of the new estimates is provided in Annex A. Although the precise conversion of the linking errors on the logit scale to the PISA scale is not reported by PISA, the scale conversion given for 'unreported gender' is used to produce the PISA Scale values in Table 5.5, where PISAScale=Logit/ 1.1002 * 100. Without more precise information, which remains unavailable at the time of this analysis, the values reported on the PISA Scale using the revised method in Table 5.5 should be used with caution.

Table 5.5: Various estimates of PISA international linking error, PISA 2000 to PISA 2009

	Published Values	Revised method			
	PISA Scale (International)	Logit Scale (International)	Logit Scale (Ireland)	PISA Scale* (International)	PISA Scale* (Ireland)
PISA 2000 to 2003	4.740	0.05950	0.14340	5.41	13.03
PISA 2003 to 2006	5.307	0.06320	0.17820	5.75	16.2
PISA 2006 to 2009	4.069	0.04500	0.14650	4.09	13.31

*Note. Interpret these estimates with caution, as they are based on unconfirmed scale adjustments.

Although the 2009 results are reported on the PISA 2000 scale, the linkage between the two cycles is based on a chain of linkages, 2000 to 2003, 2003 to 2006 and 2006 to 2009. Although the scaling procedure treats the item parameters as 'fixed and known' they are in fact estimated from a sample with an unknown sampling error. Thus, when the link is established from 2009 to 2006, there is an additional error that must be considered, because the location of the 2006 scale is somewhat imprecise *vis a vis* the 2003 scale, and likewise for 2003 back to 2000. As a result, the linking errors from cycle to cycle must be added across all links between two endpoints in the chain that are to be compared, such as 2000 to 2009. Assuming that the linkages are stochastically independent, this can be done by taking the square root of the sum of the squared errors, which produces the values in the bottom row of Table 5.5. However, these values should be taken as approximations with an unknown bias. The case of chained equating in PISA uses, for example, the same sample of items and students from PISA 2003 to estimate both the PISA 2000-2003 linkage as well as the PISA 2003-2006 linkage (and similarly for PISA 2006 with respect to PISA 2003 and PISA 2009). As a result there is a dependency in the linkage estimates between any internal link in a linking chain. This issue is described in detail by Zeng, Hanson & Kolen (1994). Accordingly, the estimates in Table 5.6 reports the linking error associated with the two major chains in PISA: 2000 to 2006 and 2003 to 2009. The methodology used to produce these estimates is explained in Annex A.

Table 5.6: Estimates of link error for PISA 2000-2006 and 2003-2009

Linking Chain	International		Ireland	
	Logit Scale	PISA Scale	Logit Scale	PISA Scale
2000-2003-2006	0.0724	6.58	0.13860	12.6
2003-2006-2009	0.0649	5.9	0.10840	9.85
2000-2003-2006-2009*	0.08670	7.88	0.191	17.36

*Estimates for this chain assume the error covariance between 2000 and 2009 is equivalent to the average of the two interim error covariances. This may result in slight underestimation of the chained linking error.

Note that the estimates for both year-to-year and chained linkages are substantially higher for Ireland than for the International sample. This larger imprecision reflects the general instability of the items in Ireland, as modelled by the Rasch model, and the larger sampling error in the Irish samples.

Although some imprecision is unavoidable due to the lack of necessary information on the scale conversion, these values are correct to an order of magnitude. The disagreement and absence of a perfect correlation with the published errors illustrates the point made by Gebhart and Adams (2007) that the issue of estimating linking error is not yet resolved with PISA. Using a conservative linking error of 9 scale points (approximately the average of the four estimates in Table 5.5), when the results of PISA 2009 are compared to PISA 2000, the combined linking, measurement and sampling error is $(7.88^2 + 3.2^2 + 2.96^2)^{0.5}$ or approximately 9.

The coefficient of variation (CV) is a statistic commonly used to gauge the interpretability of ratio-property statistics, such as differences, absolute counts, and so on. It is calculated as the standard error of the statistic divided by its estimate. Given the estimated error of the difference between Irish PISA performance from 2000 to 2009, the CV equals 9/31=29%. Following standard rules of data quality interpretation for coefficients of variation in Table 5.7, the difference in average PISA reading performance of 31 points is at the upper limit of acceptability.

Table 5.7: Coefficient of variation interpretive guidelines*

Coefficient of variation	Quality of estimate
Less than 20 percent	Very good
From 20 to 29 percent	Acceptable
From 30 to 39 percent	Use with caution
40 percent or more	Too unreliable to be published

*Note. These guidelines are approximate and should not be interepreted as rigid thresholds.

6. Summary of findings

All the OECD standards in terms of coverage, school and within-school exclusion and participation rate requirements were met by Ireland, for all PISA cycles.

A review of the PISA 2009 Quality Monitoring Report revealed some signs of student disengagement with PISA which could explain in part a drop in performance. Further analyses of "not reached" items or incomplete test booklets would be helpful to evaluate the extent to which this may have contributed to a decline in estimated mean performance.

The review made by the SCG included a close examination of the target populations, exclusion rates, sampling and estimation processes that were implemented, for both 2000 and 2009. The assessment shows that the 2000 and 2009 samples are valid and comparable to the corresponding populations from which they were drawn. Changes observed in the population characteristics between 2000 and 2009 are assumed to be a reflection of true changes in the population composition. Nothing was found in the sampling and weighting methodology that could have created these changes.

This review also included a verification of the process implemented to select the 2009 PISA sample after the selection of the ICCS sample (controlling for the overlap), as well as the changes in the 2009 stratification strategy (i.e., introducing the "% fee waiver" in stratification). The SCG concludes that these changes had no impact on the comparability of the 2000 and 2009 samples.

The SCG examined seven schools with an average score in reading of more than 100 points below the national average to verify if there were any issues with the sampling of these schools, the weights or weight adjustments. Nothing unusual was found in the sampling or weighing processes that would explain these low achievement scores.

A comparison of achievement in the 39 schools that participated in both 2000 and 2009 PISA cycle have on average a similar decline to the one observed nationally. Nothing in the sampling or weighting procedures was found that could have resulted in changes in performance.

The assessment made of the sampling methodology, mainly in 2000 and 2009, fails to find a link between the sampling and weighing methodology applied, including changes made in the sampling strategy between the 2000 and 2009, and the decline observed in the estimates of reading achievement.

The scaling of Irish PISA response data is consistent with international practice, and there is no indication that linking items are biased for or against Ireland. The content of the linking assessment reflects the balance of content areas and domains initially set for the PISA reading assessment. However, the small number of items used for linking, as well as their sparse distribution across booklets, produces a statistically unstable link between cycles.

Although PISA scaling methods do not introduce any bias in student reading estimates, international calibration of item parameters appears to be misrepresenting student performance in Ireland. Items unique to the 2009 assessment have international parameter estimates that result in estimates of student performance that are associated with lower item performance than actually achieved by Irish students.

The question of using national item parameters raises the issue of the intent of PISA and how to best achieve its goals. The national parameters estimated for Ireland are less stable than the international parameters; the degree of disagreement between Irish and international parameters is of similar magnitude to the disagreement between Irish estimates in adjacent

years. Based on these results, there may be little to gain for using national parameters in Ireland, particularly given the cost of losing international comparability of estimated results.

Notwithstanding the misrepresentation of performance by the international parameters, student performance in Ireland *is* consistently decreasing over time. Decreases in student performance on an item-by-item basis for common test items are also accompanied by increasing proportions of missing responses and an increased number of not-reached items with each cycle. These results suggest that systemic factor related to student performance are likely influencing PISA results. These may or may not be the result of declining student proficiency, but it is important to consider alternate explanations, such as student engagement and changing conditions of test administration.

While it appears that PISA can be used to identify trends over time, it does not appear that the quantification of these trends can be reported with much accuracy. The aforementioned issues with the generalization of international parameters to Ireland as well as ambiguity about the stability of the PISA international trend estimates questions the straightforward interpretations of arithmetic differences in performance over time. It is unlikely that a methodological or statistical correction will be able to adjust for these errors. Additional research is required to identify the reasonable limits of interpretation for PISA trends.

7. Recommendations

Although Ireland met the requirements in terms of participation, the student participation rate has been well below the average over all participating countries in PISA. Improving participation rates at the school and student level would reduce the risk of bias. Having make-up sessions for students who did not take part in the initial test administration could be implemented. More analyses on student non-response would also be beneficial, for example a follow-up study for the non-participating students to have a better understanding of who these students are (if they are different from the other respondents in terms of other characteristics) and why they did not participate.

More analyses should be made on schools that participated in more than one cycle (2003 and 2009, 2006 and 2009) to compare their achievement between cycles, not only in reading but also in mathematics and science and also to see if there are any changes in their characteristics that could explain the changes in achievement. Some of the schools that participated in both 2000 and 2009 saw a decrease in achievement of 100 points. A closer examination of these schools and their characteristics should be done.

In order to effectively communicate trend results in PISA over time, it will likely be necessary to estimate trends using national item parameters. Doing this will require a greater transfer of information about PISA procedures and process documentation to participating countries than currently exists. While it may be possible for the PISA Consortium to produce and distribute these results, the current experience suggests that will be advantageous for countries to receive well-documented command files for data preparation and statistical analysis in order to perform replications of scaling methods under a variety of different assumptions.

Despite already leading international assessments in the communication of linking error, greater consistency and transparency are required in incorporating this error in reporting trends. Linking error is (generally) the largest source of random error in PISA and should be considered whenever making comparisons of performance to any static benchmarks or historical results.

A key finding of this review for interpetation of the PISA 2009 results is that student performance in Ireland is decreasing, but the amount of decrease cannot be precisely quantified. The estimated drop of 31 points between PISA and PISA 2009 should be interpreted with caution. Although all evidence suggests that this drop is not purely due to chance, it should be noted that there is a conceptual difference between being non-random and being accurate. Tests of statistical significance determine if a value is statistically not equal to zero. If a value is significantly different from 0, it may still have a great deal of inaccuracy. The sampling confidence interval should be used to nuance any substantive

interpretation of the statistics. For example, the 95% confidence interval for the change in average performance between 2000 and 2009 spans 30 points on the PISA scale (1.96*7.88 points below the mean + 1.96*7.88 points above the mean). Given this wide range and the known effects of regression to the mean, any interpretation of the results used as the justification for change in action should be corroborated with independent data sources.

8. Acknowledgement

The SCG would like to thank the ERC for their availability to answer questions and for providing the necessary documents and files required to do this work.

9. References

Adams, R.J. & Wu, M.L. (2002). *PISA 2000 technical report*, OECD, Paris.

Cosgrove, J., Shiel, G., Archer, P, Perkins, R. (2010). *A comparison of performance in Ireland on PISA 2000 and PISA 2009: A preliminary Report to the Department of Education and Skills*. Internal document of the Educational Research Centre.

Gebhardt, E. & Adams, R.J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement, 8(3),* pp. 305-322.

Keyfitz, N. (1951). Sampling With Probabilities Proportional to Size: Adjustment for Changes in Probabilities, *Journal of the American Statistical Association*, Vol. 46, pp. 105-109.

Mazzeo, J., & Davier, M. von (2008). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results*. Retrieved 14/09/2010 from https://edsurveys.rti.org/PISA/documents/MazzeoPISA_Test_DesignReview_6_1_09.pdf

Monseur, C., and Berezner A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement, 8(3),* pp. 323-335.

OECD (2005). *Technical Report for the OECD Programme for International Student Assessment 2003*, OECD, Paris.

OECD (2009), *Technical Report for the OECD Programme for International Student Assessment 2006*, OECD, Paris.

OECD (2010). *PISA 2009 Main Study Data Adjudication Report*. 29th Meeting of the PISA Governing Board, April 2010, Copenhagen, Denmark.

Zeng, L., Hanson, B. A., & Kolen, M. J. (1994). Standard errors of a chain of linear equatings. *Applied Psychological Measurement, 18,* pp. 369-378.

Annex A

Derivation of linkage errors for weighted, clustered, chain-equating structures

The documentation provided by the PISA Consortium describing the calculation of linking errors for PISA is inadequate to reproduce accurate estimates of the linking error required to make reasonable inferences regarding change in country performance between PISA 2000 and PISA 2009. The methodology described in, for example, the PISA 2006 Technical Report (OECD, 2009, pp 158-159) appears to have internal notational inconsistencies inasmuch as they are inconsistent with the methodology described in Monseur and Berezner (2007). One expected difference between the Monseur and Berezner methodology should be the application of maximum point values for each item as minimum-error-variance weights for the calculation of variance components described in Monseur and Berezner. This modification is recommended by Monseur and Berezner and is also suggested in the Technical Report: "As such, items should be weighted by their maximum possible score when estimating the equating error" (p 159). However, the formulae provided in the Technical Report do not consistently reflect such an approach.

The derivations in this appendix describe a methodology that is consistent with the approach described by Monseur and Berezner (2007) with the application of maximum score weighting for each item. In addition, this methodology is applied for the calculation of error covariances between non-adjacent years, which affects the magnitude of error for linking chains, such as the multi-step link from PISA 2003 to PISA 2006 to PISA 2009. Finally, all sources of error are combined to estimate the total chained linking error for nonadjacent PISA cycles.

As with previous methods, the estimation process begins with calculating the difference in item difficulty, c_i, for each item, i.

$$c_i = \hat{\delta}_i^1 - \hat{\delta}_i^0 \qquad (1)$$

where $\hat{\delta}_i^0$ is the estimate of item difficulty in the preceding Cycle and $\hat{\delta}_i^0$ is the estimate of difficulty in the following Cycle. Each $\hat{\delta}_i^0$ is centered around each Cycle's average such that

$\sum_{i}^{n} m_i \hat{\delta}_i = 0$, using the value, *m*, the maximum score point for an item, as a weight. The index, *i* refers to individual items. Each item is also nested with one of *K* item units, which are indexed by *j*.

The description of method provided in Monseur and Berezner uses simple sums of squared deviations and raw totals to calculate the mean squared deviations. The unweighted formulae,

$$SS_w = \sum_{j}^{K}\sum_{i}^{n}\left(c_{ij} - \overline{c_{\bullet j}}\right)^2$$

$$SS_B = \sum_{j}^{K}\left(c_{\bullet j} - \overline{c}\right)^2$$

are modified here using the item weights:

$$SS_w = \sum_{j}^{K}\sum_{i}^{n} m_i\left(c_{ij} - \overline{c_{\bullet j}}\right)^2 \quad \text{, and} \qquad (2)$$

$$SS_B = \sum_{j}^{K}\left[\left(\sum_{i} m_{ij}\right)\left(c_{\bullet j} - \overline{c}\right)^2\right] \quad . \qquad (3)$$

To calculate Mean Square components, the following unbiased estimator was used:

$$MS = \frac{V_1}{V_1^2 - V_2} SS \qquad (4)$$

where , $V_{1w} = V_{1B} = \sum_{j}\sum_{i} m_{ij}$, $V_{2W} = \sum_{j}\sum_{i} m_{ij}^2$ and $V_{2B} = \sum_{j}\left(\sum_{i} m_{ij}\right)^2$.

Following Monseur and Berezner, the unbiased estimator of the between cluster variance is $\sigma_B^2 = \frac{\left(MS_B - MS_W\right)}{\bar{n}}$, where $\bar{n} = \frac{V_{1B}}{K}$, and the total error variance for the linkage is

$$\sigma_{linkage}^2 = \frac{\sigma_B^2}{K} - \frac{\sigma_W^2}{V_{1W}}.$$

As with Berezner and Monseur, we assume the link items to represent an infinite population, an assumption further warranted by the fact that almost an equivalent amount of the measurement information used for each student is taken from corollary variables and other content domains as is taken from the actual test items within each domain.

This methodology is also used to calculate the covariance errors between nonadjacent cycles, with the small modification that, instead of using squared deviations in Equations (2) and (3), the crossproducts are used, as an application of the general formulation of the Sums of Squares and Cross Product matrix:

$$SSCP_{xy} = \sum\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)$$

For example,

$$SS_{B2000\_2006} = \sum_{j}^{K}\left[\left(\sum_{i} m_{ij}\right)\left(c_{\bullet j2000\_2003} - \overline{c_{02000\_2003}}\right)\left(c_{\bullet j2003\_2006} - \overline{c_{2003\_2006}}\right)\right].$$

Using the estimates of the SSCP matrix for the two-step chained linkages (i.e., 2000 to 2006 and 2003 to 2009), the estimates of chained linking error are estimated as:

$$\sigma^2_{chained} = \sigma^2_{linkageX} + \sigma^2_{linkageY} + 2\left(\sigma^2_{linkageXY}\right). \qquad (5).$$