The Irish Journal of Education, 2015, *xl*, pp. 29-44.

CHANGES IN ACHIEVEMENT IN PISA FROM 2000 TO 2009 IN IRELAND: BEYOND THE TEST SCORES

Jude Cosgrove
Educational Research Centre
St Patrick's College, Dublin

The results for PISA 2009 are revisited, focusing on the large decline in the scores of Irish students in reading and mathematics achievement. Findings concerning two aspects of PISA are reviewed: (a) the PISA test design and (b) the methods used to scale PISA achievement scores and to link achievement across cycles. Analyses suggest that changes across cycles in the relative weightings of the item formats and cognitive processes assessed in the PISA tests had unintended consequences for estimating trends, at least in the case of Ireland. Questions are raised about scaling methodology and PISA's methods for reporting the accuracy of estimates of change, though the extent to which these might affect Irish students in a unique fashion is unknown. Findings are considered in conjunction with changes in the PISA population over time and a possible decline in the engagement of Irish students with the PISA tests. Finally, whether and how future cycles of PISA might address some of the issues raised is considered.

When first published, the results for PISA 2009 reading, and to a lesser extent for mathematics, attracted considerable media attention and commentary in Ireland. The *Irish Times* headlined the results as 'shattering the myth of a world-class education system' (December 8, 2010), while the *Irish Independent* commented that: '[T]here was shock last year when it emerged there was a fall in reading and maths scores for Irish students in the PISA' (April 3, 2012). *Education Matters* described the results as 'an urgent call to action' (December 14, 2010). These comments were a reaction to the finding that Ireland's mean reading score on PISA showed the largest decline since 2000 (31 score points, or close to one-third of an international standard deviation) across the 38 countries for which results could be compared; mathematics showed the second-largest decline since 2003 (16 points, or one-sixth of an international standard deviation) across the 39 countries that could be compared. Achievement in science remained stable (OECD, 2010b). These results were not supported by any evidence of a decline in educational standards in Ireland. Indeed, the results of the most recent cycle of PISA (PISA 2012) show that Ireland's mean scores in mathematics and reading

were about the same as they had been in 2003 and 2006, while the results for science in 2012 were significantly higher than they had been in 2006 and 2009 (Perkins, Shiel, Merriman, Cosgrove, & Moran, 2013).

Even before the results of PISA 2012 had become available, the need was recognised to attempt to account for the 2009 findings and to challenge the kneejerk reaction of the media. First, the Department of Education and Skills (Ireland) decided to seek input from independent international experts, leading to two comprehensive reviews (Cartwright, 2011; LaRoche & Cartwright, 2010). Second, staff at the Educational Research Centre undertook additional analyses to identify possible explanations of the decrease in scores of Irish students, particularly in reading (Cosgrove, 2011; Cosgrove, & Moran, 2011; Cosgrove, Shiel, Archer, & Perkins, 2010; Shiel, Moran, Cosgrove, & Perkins, 2010). Analyses, the results of which are presented in this paper, focused on features of the PISA test design and methods used to estimate change over time in achievement. Third, the possible effects of demographic change in the PISA cohort in Ireland during the lifetime of PISA were considered (Perkins, Cosgrove, Moran, & Shiel, 2012). Finally, changes in student engagement with PISA tasks were investigated by scrutinizing student performance on the first and final quarters of tests to determine if performance differed on the two segments of the tests (Cosgrove & Cartwright, 2014).

This paper is organized into four main sections. First, whether and how aspects of the PISA test design, and changes to it between 2000 and 2009, are related to student achievement is considered. Second, aspects of PISA's approach to producing achievement scores and to linking scores across cycles are reviewed. Third, the findings of studies that consider the possible effects of demographic changes on student test scores are described. Finally, the results of analyses that were carried out to explore the possibility that changes in students' engagement with PISA tests could have impacted on their performance are considered. The paper concludes by summarizing key observations and making some recommendations for international test design and scaling. In doing so, consideration is given to the extent to which more recent and current cycles of PISA reflect these[1].

[1] Aims, design and results of PISA are not considered here. Readers are referred to reports on PISA 2009 (OECD, 2010a; OECD, 2010b; OECD, 2010c; OECD, 2010d; OECD, 2010e; in particular OECD, 2010b), the PISA 2009 technical report (OECD, 2011), and the PISA 2009 assessment framework (OECD, 2009). Additional international reports are available at *www.oecd.org/pisa*, while national reports on PISA are available for Ireland at *www.erc.ie/pisa*.

FEATURES OF THE PISA TEST WITHIN AND ACROSS CYCLES, AND THEIR ASSOCIATION WITH ACHIEVEMENT

This section considers the extent to which the PISA tests vary across cycles in their design features, and whether or not these variations are related to achievement. Table 1 shows the percentages of items of differing format used in PISA 2000, 2003, 2006, and 2009. Although the fact that the number of items in a test changes from major to minor domains makes comparisons difficult, two general patterns are evident. First, there is a decrease in the percentage of written response items in all domains (in reading, this is more evident since 2003), and an increase in the percentage of complex multiple-choice items. Second, changes in the representation of regular multiple-choice items vary from domain to domain: the percentage of these items increased across cycles in mathematics and decreased in science, with reading showing a decrease in 2003 and 2006, and an increase in 2009.

Table 1
Representation of Item Response Types, by PISA Domain and Cycle

Domain/Response Type	Distribution of items (%)			
	2000	2003	2006	2009
Reading				
Written response	54.6	64.3	64.3	52.9
Complex multiple-choice	4.4	3.6	3.6	8.1
Multiple-choice	41.0	32.1	32.2	39.0
Mathematics				
Written response	67.3	66.1	56.3	54.0
Complex multiple-choice	12.9	13.4	18.7	20.2
Multiple-choice	19.9	20.5	24.9	25.8
Science				
Written response	41.6	41.1	35.6	34.1
Complex multiple-choice	17.9	20.7	29.7	31.9
Multiple-choice	40.5	38.3	34.7	34.0

Source: Cartwright, 2011, Table 4.

The representation of PISA cognitive subscales by domain also varies across cycles (Table 2). Since the representation of subscales is not inherently part of the PISA design until a scale is established as a major domain (in the case of reading, this was in 2000; for mathematics, it was in 2003; and for science, in 2006), figures for mathematics prior to 2003 and for science prior to 2006 are not considered here.

In reading, the changes relate primarily to a decrease in Access and Retrieve items with a corresponding increase in Integrate and Interpret items. In mathematics, changes in the representation of subscales primarily involve an increase in Quantity and decreases in Space and Shape and Uncertainty. No clear pattern is evident in science.

Table 2
Representation of Cognitive Subscales, by PISA Domain and Cycle

Domain/Subscale	Distribution of items (%)			
	2000	2003	2006	2009
Reading				
Access and retrieve	27.7	25.0	24.9	22.8
Integrate and interpret	49.3	49.9	50.1	52.1
Reflect and evaluate	23.0	25.1	25.1	25.1
Mathematics				
Change and relationships		24.3	25.1	25.7
Quantity		26.5	26.9	31.5
Space and shape		25.2	24.9	22.8
Uncertainty		24.0	23.0	20.0
Science				
Explaining phenomena scientifically			47.5	41.5
Identifying scientific issues			22.8	24.4
Using scientific evidence			29.7	34.1

Source: Cartwright, 2011, Table 5.

Cartwright (2011) has shown that these aspects of the test design interact with students' response patterns in a manner that is consistent with changes in overall Irish performance on PISA. The results in Table 3 summarize student performance (expressed as percent correct) on each of the item response types for each domain across cycles. On the reading assessment, between 2000 and 2009, performance on both regular multiple-choice and complex multiple-choice items declined substantially. On the mathematics assessment, between 2003 and 2009, performance on written response items declined markedly, while performance on both regular and complex multiple-choice items remained stable. There are no such marked changes on the science assessment.

Table 3
Difficulty of Item Response Types for Students in Ireland, by PISA Domain and Cycle

Domain/Response Type	Percent Correct			
	2000	2003	2006	2009
Reading				
Written response	61.6	60.8	59.3	60.4
Complex multiple-choice	62.7	61.8	57.1	43.2
Multiple-choice	72.1	72.2	71.7	63.4
Mathematics				
Written response	43.4	50.0	46.5	37.9
Complex multiple-choice	37.9	48.4	43.8	49.8
Multiple-choice	64.7	55.6	56.0	57.9
Science				
Written response	46.3	46.9	45.7	48.8
Complex multiple-choice	53.4	51.7	60.0	55.2
Multiple-choice	56.6	57.3	61.5	59.6

Source: Cartwright, 2011, Table 6.

Changes over time were also recorded in the performance of Irish students on cognitive subscales for all domains (Table 4). The changes may be partly attributable to other factors, including item format. In reading, where reduced item sets were administered in 2003 and 2006, and expanded sets in 2000 and 2009 (albeit containing the 2003 and 2006 item sets), there was a noticeable improvement between 2006 and 2009 in performance on Access and Retrieve items; however, this was offset by the decrease in representation of these items, and the more gradual performance decline on Integrate and Interpret items. In mathematics, where the same item sets were administered in 2006 and 2009, and an expanded set that included the 2006 and 2009 items was administered in 2003, performance tended to decrease, most markedly for Space and Shape; Uncertainty is the only subscale where performance remained relatively constant. Increasing representation of Quantity items moderates the more strongly negative influence of Space and Shape. There are no clear patterns over time in the changes in the difficulty of science items by subscale. However, it is notable that performance on science stayed fairly constant in the presence of large negative changes in mathematics and reading.

Table 4
Difficulty of Item Cognitive Subscales for Students in Ireland, by PISA Domain and Cycle

Domain/Subscale	Percent Correct			
	2000	2003	2006	2009
Reading				
Access and retrieve	69.0	58.1	54.2	70.4
Integrate and interpret	67.8	69.3	68.4	58.0
Reflect and evaluate	58.2	61.2	61.6	55.5
Mathematics				
Change and relationships		52.3	52.0	45.4
Quantity		58.3	55.6	51.5
Space and shape		43.1	37.9	32.1
Uncertainty		49.7	47.3	51.3
Science				
Explaining phenomena scientifically			56.5	55.0
Identifying scientific issues			57.8	56.0
Using scientific evidence			51.9	52.8

Source: Cartwright, 2011, Table 7.

Table 5 presents the results of a decomposition of variance of the individual scored item responses into school, student, item response type (coded response, multiple-choice, or complex multiple-choice), and item cognitive subscale components. Typically, any variance within students' item responses is interpreted as random measurement error and is ignored during secondary analysis of results. However, the data in Table 5 indicate that the variance within each student, including the item response type, subscale and unexplained components, accounts for a large proportion of the total variance. This reflects the diversity of the items to which students were required to respond. More traditional educational assessments that test a narrow set of skills using a single item format tend to have lower proportions of variance within students' responses. The pattern of note in the table is that the components attributable to the PISA design (response type and subscale), though low overall, tend to vary substantially. In PISA 2009, for example, the percentage of variance in item scores attributable to item response type was more than double the percentage attributable to schools in the case of both mathematics and reading.

Table 5
Percentages of Variance in Scored Item Responses Attributable to Various Components, Ireland

Domain/Year	Variance accounted (%)				
	School	Student	Response type*	Subscale	Unexplained
Reading					
2000	2.8	12.4	1.3	1.0	82.5
2003	3.6	13.4	0.8	0.4	81.8
2006	3.2	15.8	0.9	0.9	79.1
2009	3.3	12.0	6.6	3.2	75.0
Mathematics					
2000					
2003	2.5	11.3	0.7	1.8	83.7
2006	2.5	10.7	2.6	3.1	81.2
2009	2.0	11.6	4.6	2.8	79.0
Science					
2000					
2003					
2006	2.3	11.9	3.0	0.0	82.8
2009	2.9	11.8	1.2	0.1	84.0

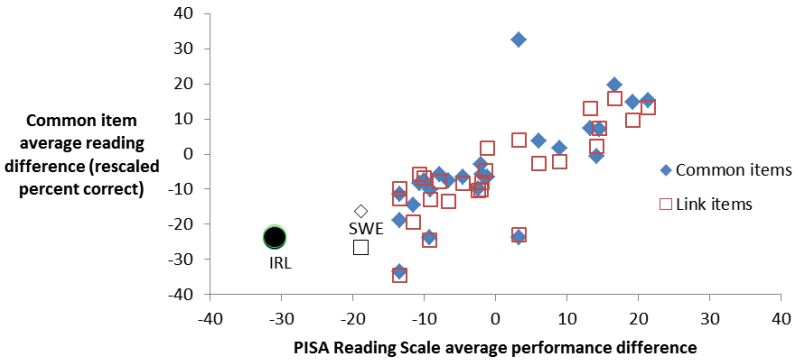
Source: Cartwright, 2011, Table 3.
*Written response, complex multiple-choice, regular multiple-choice.

Table 5 shows that fluctuations in design elements of PISA (i.e., item type and subscale), in particular item response type for reading and mathematics, are related to student performance. However, documentation on the PISA scaling methods indicates that interactions between performance and design are not taken into account in the estimation of student proficiency (OECD, 2011). As a result, a key factor influencing the reported changes in performance over time, at least in the context of Ireland, appears to be the change in operational definition of the domains being tested (i.e., how the assessment frameworks are translated into test content and format). Unfortunately, because there is no control sample that has been assessed over the same period using constant or unmodified operationalizations of the PISA assessment domains, it is not possible to definitively state that the reported changes over time are entirely artefacts of the changes in design.

PISA'S APPROACH TO ESTIMATING CHANGES IN ACHIEVEMENT

As a starting point for considering PISA's methods of estimating change over time, it is useful to illustrate the correspondence between changes in percent correct scores on reading link items and changes in PISA reading scores between 2000 and 2009 as reported by the OECD (2010e). If the scaling and linking methods are unbiased, one would expect a close correspondence between these two estimates. Figure 1 plots the changes in percent correct and PISA achievement scores for reading between PISA 2000 and PISA 2009. In producing this figure, Cartwright (2011) estimated changes in percent correct both for link items (used in 2000, 2003, 2006, and 2009) and for common items (used in 2000 and 2009 only). There is, with some exceptions, a close correspondence between the two estimates of change.

Figure 1
Comparison of Differences in Average Item Performance in Reading to Reported Differences in PISA Reading Proficiency for Countries in the PISA Population Between 2000 and 2009



Source: Cartwright, 2011, Figure 4.
Ireland is represented by the black dot, which is the location of both common and link items.
Sweden is represented by the white markers.

A decline in the percentage of correct responses in Ireland was recorded for Irish students, but the decline is not as large as PISA scaled scores (used in international and national reporting of results) would indicate. For example, Sweden (marked in white) has the same change in the percent correct on link

items as Ireland, yet its PISA reading score decline is only 19 points (compared to 31 points in Ireland). Why, therefore, is there not a closer correspondence between percent correct estimates and scaled scores?

Cartwright (2011) has identified two issues concerning the Rasch model used to produce PISA achievement scores that may affect estimates of change. First, item discrimination is fixed, i.e., items are constrained to be equivalent in terms of the strength of their relationship with proficiency across countries and sub-groups within countries. Cartwright has demonstrated that this constraint is inappropriate for both the OECD on average and Ireland, and that proper modelling of the PISA items would require an item response model that would allow item discrimination to vary (see also Mazzeo & von Davier, 2008). Thus, the issue is considered a general one, not one confined to Ireland.

A further issue with the Rasch model used to produce achievement scores that may affect estimates of change is that the items are assigned parameters that are calculated on the basis of an artificial population (the PISA calibration sample, which consists of a random sub-sample of the same number of students from each participating OECD country). This issue becomes problematic if the Rasch model represents a systematic misfit to a specific country (rather than misfitting in a non-systematic or random way). To address this issue, Cartwright (2011) conducted a re-calibration of achievement for Ireland and internationally in reading, mathematics, and science on the basis of national item difficulties. Results indicate that PISA reading data are more sensitive to model specification and item calibration than mathematics or science. The relative sensitivity of the PISA reading assessment to model specification may be due to the smaller number of reading link items used to estimate change, responses to individual items being more dependent on the passage on which they are based than in science or mathematics (see Monseur, 2009), or to some other aspect of the PISA design.

A third issue, that may or may not be specific to the case of Ireland, has been documented by LaRoche and Cartwright (2010) who found a systematic model misfit for Ireland for reading in PISA 2009, which likely resulted in the reported PISA reading score for students in that year being an underestimate of achievement. The misfit appears to be due to the non-equivalence of new and link reading items administered in PISA 2009; the assumption of the scaling model is, of course, that they are equivalent (see OECD, 2011, Chapters 9 and 12). However, since the analyses were limited

to Ireland, we do not know the extent to which model misfit may have affected score estimates for other countries.

Also relevant to how linking may be problematic in estimating the significance of change in administering PISA is the manner in which the link error is computed (Gebhardt & Adams, 2007; LaRoche & Cartwright, 2010). The method used to link PISA 2000 and PISA 2003 underestimated the error (Monseur & Berezner, 2007) and was subsequently revised (OECD, 2005). However, details on the precise derivation of the link errors are lacking (see OECD, 2011, Chapter 12). LaRoche and Cartwright (2010), having explored alternative methods to compute linking error, concluded that the OECD (2010e; 2011) consistently underestimated the magnitude of the error between 2000 and 2003, 2003 and 2006, and 2006 and 2009. If the OECD had estimated changes in achievement using standard errors that were larger, fewer significant differences would have been found. This issue is compounded by the fact that PISA uses chained equating. That is, the same sample of items and students from PISA 2003 is used 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. For these reasons, Cartwright (2011) is highly critical of the manner in which the OECD has represented changes in achievement (particularly data representations such as in Figure V.2.1 in OECD, 2010e), which does not take into account the issues relating to the estimation of the link error, or the complexities underlying the trend estimates.

DEMOGRAPHIC CHANGES IN THE PISA COHORT IN IRELAND ACROSS CYCLES

In a discussion of demographic changes, Cosgrove and Cartwright (2014) (also described in detail in Perkins et al., 2012) note that chief among these is an increase in the immigrant population that took part in PISA (which was the second largest increase across OECD countries since 2000). Moreover, the immigrant population in 2009 was less socioeconomically advantaged than in 2000. Policy changes over the past decade concerning retention rates and the inclusion of students with special educational needs in mainstream schools have also had a noticeable impact on the composition of the PISA student samples. The increase of students taking the optional Grade 10 (Transition Year) programme (from 16% to 24%) is likely to reflect both the increased availability of this programme and the desire of some students to

stay longer in school in the context of shrinking job opportunities at the time of the PISA 2009 administration. Since these demographic changes did not occur in isolation from each other, it is not possible to estimate their effects on the PISA scores.

CHANGES IN THE ENGAGEMENT OF IRISH STUDENTS WITH THE PISA TESTS ACROSS CYCLES

Cosgrove and Cartwright (2014), in a study of student engagement with the PISA tests, analysed response patterns of students across cycles by comparing their performance on the same sets of items when presented to students in the first and last quarters of their test booklets. The declines in percent correct in reading and mathematics in the last quarter of booklets in later cycles were due largely to an increase in the rate of students skipping items, as opposed to attempting items and getting them incorrect, strongly suggesting a marked decline in students' engagement. One would have observed a global increase in the percent of incorrect responses had there been a decline in proficiency as opposed to student engagement. Indeed, Borghans and Schils (2011) argue that 'test motivation' or engagement is conceptually and empirically distinct from ability or proficiency. Unfortunately, PISA conflates the two (as do other large-scale assessments of achievement). Whether, how, and to what extent changes in student engagement are linked with demographic and other changes is difficult to determine.

CONCLUSION

This paper considered some potential reasons for the reported declines in Ireland's reading and mathematics scores in PISA 2009, focusing on two issues that may be considered as lying inside the workings of PISA that are fundamental to its design and methodology: changes to the content of the PISA test, and PISA's methods for estimating change. Other aspects of PISA that are ostensibly outside PISA's design and methodologies are changes in the PISA population in Ireland over time, and changes in how students engaged with the PISA tests.

The task of disentangling methodological issues from ones which indicate substantive changes in proficiency is extremely complex. It is unlikely, however, that any country experienced such a confluence of confounding factors affecting the interpretation of performance trends as Ireland did in

2009. In this respect, the case study of Ireland is useful in that the overall magnitude of the combined effect of these factors stimulated a detailed examination that may not have been undertaken if their random effects had reached a zero sum.

The evidence reviewed in this paper showed that there have been marked changes in the content and structure of the PISA tests in all domains across cycles and that these are highly likely to have unintended consequences for the estimation of trends. Not only are Irish students' response patterns idiosyncratic among PISA countries, there is evidence of a general decline in engagement in the reading and mathematics tests in 2009 (Cosgrove, 2011). It is thus reasonable to conclude that changes in the design of the PISA tests and changes in the engagement of Irish students with the PISA tests over time have interacted with one another giving rise to unfavourable consequences for reading and mathematics, but oddly, not for science. Given that the computer-based assessment used in PISA 2012, and the assessment planned for 2015, capture student response latencies (or time on task), it would seem worthwhile to explore how these data could add to our understanding of what the PISA scores mean, and whether engagement and proficiency might be usefully and meaningfully distinguished from one another. Arguably, one is as important as the other, and each may suggest a distinct set of policy responses (see, e.g., Eklöf, 2007; van Barnevald, Pharand, Ruberto, & Haggarty, 2013).

The assumption that the link and new items were psychometrically equivalent is not upheld in reading in 2009 in Ireland, though the extent to which this is an issue for other countries has yet to be explored (LaRoche & Cartwright, 2010). Establishing the 2000-2009 reading link was the first time that this assumption was relied on. The equivalence of new and link items becomes even more important with further development of computer-based assessment in 2015 and beyond, where the assumption of equivalence may be even more difficult to satisfy across two assessment modes (paper-based and computer-based). It would seem desirable to avoid a 'two-tier PISA', whereby the results of countries which are not in a position to transition to computer-based assessment are not directly comparable to countries that are able to implement PISA electronically.

A further aspect of the PISA scaling methodology that has been identified previously has also been raised in this paper, i.e., the method used to compute the linking error, which, in LaRoche and Cartwright's (2010) view, is underestimated. Indeed, Gebhardt & Adams (2007) comment that 'no

consensus has been reached about [estimation of linking errors]' (p. 309), and this is evidenced by the fact that the linking error in the IEA Progress in International Reading Literacy Study (PIRLS) and Trends in Mathematics and Science Study (TIMSS) is based on a different approach.

> In contrast to the PISA approach to linking error, which is based on the variance of the linking-item difficulty parameters, the TIMSS and PIRLS approach considers linking error in terms of differences in student achievement distributions due to item parameter changes. (Martin, Mullis, Foy, Brossman, & Stanco, 2012, p. 46)

This is a potentially important issue: if the link error is underestimated and countries implement policies on the basis of what they believe to be significant changes in achievement, these run the risk of being erroneous in light of other data sources and successive cycles of PISA, as the trends become more reliable with an increased number of data points.

Looking ahead to PISA 2015, two positive developments, which can be expected to result in more stable trends and more accurate estimates of student achievement, may be noted. The first is that the Rasch model will be replaced with a two-parameter model [such as the one used in the recent PIAAC (Programme for the International Assessment of Adult Competencies) study; see OECD, 2014; Yamamoto, Khorramdel, & von Davier, 2013], permitting the discrimination of items to vary across countries where appropriate. The second is an increase in the numbers of so-called 'trend' items, or test questions that will be used to estimate change in achievement over time; this is likely to result in more stable and reliable estimates of trends in achievement. Furthermore, the data gathered in all previous PISA cycles will be re-scaled, resulting in much more robust item parameters from which to scale the PISA 2015 data (OECD, 2014).

PISA as a benchmark for monitoring the relative progress of education systems continues to grow in prominence while at the same time attempting to measure change in a world where definitions of knowledge and skills are themselves changing. These two characteristics of PISA place demands on the producers of the results it generates. The findings in this paper lend themselves to three recommendations in this regard. First, the sets of items used from cycle to cycle to estimate trends in achievement should be matched not only by content area but also by item format and cognitive process. Second, the introduction of new item formats (such as responses on the basis of mini virtual experiments or simulations in PISA 2015 computer-based science) should be assessed carefully to examine their fit with (or equivalence

to) existing item formats. Third, greater consistency between international studies on the reporting of trends, in particular their measurement errors, presentation and interpretation, should be sought by stakeholders in participating countries.

REFERENCES

Borghans, L., & Schils, T. (2011). *The leaning tower of PISA: The effect of test motivation on scores in the international student assessment.* Paper presented at the EALE annual conference, Paphos, Cyprus, September 22-24.

Cartwright, F. (2011). *PISA in Ireland, 2000-2009: Factors affecting inferences about changes in student proficiency over time.* Dublin: Educational Research Centre.

Cosgrove, J. (2011). *Does student engagement explain performance on PISA? Comparisons of response patterns on the PISA tests across time.* Dublin: Educational Research Centre.

Cosgrove, J., & Cartwright, F. (2014). Changes in achievement on PISA: The case of Ireland and implications for international assessment practice. *Large-Scale Assessments in Education, 2*, 2 (doi:10.1186/2196-0739-2-2).

Cosgrove, J., & Moran, G. (2011). *Taking the PISA 2009 test in Ireland: Students' response patterns on the print and digital assessments*. Dublin: Educational Research Centre.

Cosgrove, J., Shiel, G., Archer, P., & Perkins, R. (2010). *Comparisons of performance in PISA 2000 to PISA 2009. A preliminary report to the Department of Education and Skills.* Dublin: Educational Research Centre.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*, 311-332.

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

LaRoche, S., & Cartwright, F. (2010). *Independent review of the PISA 2009 results for Ireland: Report prepared by the Educational Research Centre at the request of the Department of Education and Skills.* Dublin: Department of Education and Skills.

Martin, M.O., Mullis, I.V.S., Foy, P., Brossman, B., & Stanco, G.M. (2012). Estimating linking error in PIRLS. *IERI Monograph Series (Vol. 5): Issues and Methodologies in Large-Scale Assessments,* 35-47.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results.* Paris: OECD Education Working Papers (EDU/PISA/GB(2008)28).

Monseur, C. (2009). *Item dependency in PISA*. Paper presented at the PISA research conference, Kiel, Germany, September 14-16.

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

OECD (Organisation for Economic Co-operation and Development). (2005). *PISA 2003 technical report.* Paris: Author.

OECD. (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science.* Paris: Author.

OECD. (2010a). *PISA 2009 results: Learning to learn – Student engagement, strategies and practices (Volume III).* Paris: Author.

OECD. (2010b). *PISA 2009 results: Learning trends – Changes in student performance since 2000 (Volume V).* Paris: Author.

OECD. (2010c). *PISA 2009 results: Overcoming social background – Equity in learning opportunities and outcomes (Volume II).* Paris: Author.

OECD. (2010d). *PISA 2009 results: Resources, policies and practices (Volume IV).* Paris: Author.

OECD. (2010e). *PISA 2009 results: What students know and can do – Student performance in reading, mathematics and science (Volume I).* Paris: Author.

OECD. (2011). *PISA 2009 technical report.* Paris: Author.

OECD. (2014). *PISA field trial goals, assessment design and analysis plan for the cognitive assessment.* Unpublished PISA Governing Board Meeting Document, April. Paris: Author.

Perkins, R., Cosgrove, J., Moran, G., & Shiel, G. (2012). *PISA 2009: Results for Ireland and changes since 2000.* Dublin: Educational Research Centre.

Perkins, R., Shiel, G., Merriman, B., Cosgrove, J., & Moran, G. (2013). *Learning for life: The achievements of 15-year-olds on mathematics, reading literacy and science in PISA 2012.* Dublin: Educational Research Centre.

Shiel, G., Moran, G., Cosgrove, J., & Perkins, R. (2010). *A summary of the performance of students in Ireland on the PISA 2009 test of mathematical literacy and a comparison with performance in 2003. Report to the Department of Education and Skills.* Dublin: Educational Research Centre.

van Barneveld, C., Pharand, S.L., Ruberto, L., & Haggarty, D. (2013). Student motivation in large-scale assessments. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment in education: Theory, issues, and practice* (pp. 43-61). New York: Routledge.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Scaling PIAAC cognitive data. In OECD (Ed.), *Technical report of the survey of adult skills (PIAAC)* (Chapter 17). Paris: OECD.