Is There a 'Crisis' In Welsh Education?[1]

Gareth Rees and Chris Taylor

Introduction

The current state of Welsh education has become a matter of widespread concern in recent years. Certainly, many of those outside Wales have formed a view of Welsh education that is extremely unfavourable. For example, following a series of well-publicized clashes between representatives of the two governments, the UK Government has made clear its belief that – as with the delivery of other public services – the Welsh Government's educational provision is highly unsatisfactory.[2] Writing recently in the *Western Mail* (28 June 2014), the then UK Secretary of State for Education, Michael Gove, summarized this position:

> Wales is an object lesson in what happens when you abandon reform. Ed Miliband recently told the Welsh Labour Conference that Wales' Labour Government is 'proving to the rest of the country the difference that Labour can make'.
>
> It's certainly doing that on education. Thousands of children falling behind as a result of rigid dogma and a refusal to reform – that's the difference that Labour has made in Wales.

As we shall see, it is difficult to sustain the case that there has been an *absence* of educational reform in Wales. Rather the controversy is over the effects of *different types* of reform, with the UK Government having pursued policies in England that differ – at least in key respects – from those adopted by the Welsh Government (and, indeed, by the other devolved administrations in Scotland and Northern Ireland). Hence, the critics of Welsh Government policy emphasize that, in Wales, schools under-perform significantly, and that this is a reflection of the educational policies that have been adopted.

It is striking, however, that the assessment of educational performance (and hence of the effects of alternative policies) has taken a particular form. Firstly, the focus has been on statistical evidence of levels of educational attainment – narrowly defined – in Welsh schools. Other measures of educational performance could

1 We are very grateful for the financial support of the Honourable Society of Cymmrodorion, which made the research here possible; and for the invaluable advice and support provided throughout by two members of the Council of the Cymmrodorion, John Elliott and Ceridwen Roberts. We should also like to thank members of the Welsh Government for making data available and for commenting on earlier drafts. The analysis was improved significantly by the comments provided by an external reviewer, to whom we should like to offer warm thanks. The views expressed are those of the authors only.

2 Leighton Andrews, the former Minister for Education and Skills in the Welsh Government, provides a vivid account of what he terms 'the war on Wales' in his recently published book, *Ministering to Education* (Cardigan: Parthian, 2014).

clearly be adopted. For example, in a recent article in the *British Medical Journal*, Bonnell *et al.* make the argument (specifically in relation to schools in England) that the health and wider well-being of children and young people should be set alongside educational attainment as measures of the effectiveness of schooling, not only because of the strong positive relationship between the two sorts of outcomes, but also the importance of health and well-being in their own right.[3]

Robust measures of well-being are not readily available across the student population. However, the Millennium Cohort Study (MCS) – a panel survey of a representative sample of some 19,000 children born in the UK during 2000–2001 – does reveal that seven-year-olds in Wales score significantly higher on measures of personal well-being and positive attitudes towards learning than their peers in England; and somewhat higher on measures of well-being at school.[4] It remains to be seen, however, how far these differences are carried forward, as the MCS cohort progresses through the education system.

It is also striking that, despite the widespread perception that Welsh schools are underperforming, the majority of parents remain 'very satisfied' or 'fairly satisfied' with their children's schools. The National Survey for Wales for 2012-2013 – a survey conducted by the Welsh Government of a representative sample of almost 15,000 adults across Wales – reveals, for example, that 65 per cent of parents were 'very satisfied' with their children's primary school; and a further 27 per cent were 'fairly satisfied'. The equivalent figures for secondary schools were 50 per cent 'very satisfied' and 33 per cent 'fairly satisfied'.[5] Clearly, we should not wish to make too much of this sort of evidence, as it is open to a wide variety of interpretations. However, it casts a fresh light on the performance of schools in Wales, especially for those who favour 'parental choice' as an organising criterion for educational provision. Again, it serves to qualify the unremittingly bleak account of Welsh education that has come to dominate the public debate.

However, these sorts of 'soft' indicators have generally been ignored in this debate.[6] Rather, it is educational attainment which is deemed to provide the key measure of the effectiveness of Welsh schools; and hence the impacts of divergent policy initiatives. What this underlines, then, is that decisions as to what are the most appropriate indicators of performance reflect judgements about what is important in

3 C. Bonnell, N. Humphrey, A. Fletcher, L. Moore, R. Anderson and R. Campbell, 'Why Schools Should Promote Students' Health and Well-being', *BMJ*, 348: g3078 (2014).

4 C. Taylor, G. Rees, and R. Davies, 'Devolution and Geographies of Education: The Use of the Millennium Cohort Study for "Home International" Comparisons across the UK', *Comparative Education*, 49.3 (2013), 290–316.

5 R. Cook, J. Morrison and L. Phillips, *National Survey for Wales: Satisfaction with Education, Final Report*, Welsh Government Social Research 29/2014 (Cardiff: Welsh Government, 2014). These reported levels of parental satisfaction are broadly consistent with those recorded in Scotland through *Growing Up in Scotland*, a large-scale longitudinal survey that follows the lives of children from infancy to their teens, http://growingupinscotland.org.uk/. Unfortunately, as far as we are aware, comparable data are not available for England.

6 It is instructive, however, that the CBI in Wales has recently argued the case for considering a wider range of criteria against which to assess the performance of schools than educational attainment alone: CBI Wales, *Step Change: A New Approach for Schools in Wales* (Cardiff: CBI, 2014).

educational systems. Certainly, the predominance here of conventional measures of educational attainment reflects a particular orientation towards education's primary role, which is defined essentially in terms of the perceived positive relationships between educational attainment and economic development.

Secondly, the benchmark for judging levels of educational attainment in Wales has been comparison with those in other countries. Hence, educational performance in Wales has not been judged primarily by reference to, for example, changes in Welsh attainment levels over time; but rather by comparing levels of attainment in Wales with those in other parts of the UK – and England, in particular. Indeed, this has been extended more recently to wider international comparison, based on the Programme for International Student Assessment (PISA), a survey of the attainment of 15-year-olds in reading, mathematics and science, conducted every three years in a wide range of countries by the Organisation for Economic Co-operation and Development (OECD). Wales participated in its own right in PISA for the first time in 2006; and PISA scores have subsequently become a key element in the debate.[7] It is these *cross-national* comparisons of educational attainment, therefore, that have provided the 'evidence base' for the performance of Welsh schools and the political controversies surrounding the impacts of different policy approaches. As we shall see, however, interpreting such evidence is highly complex.

It should be emphasized, nevertheless, that there is substantial agreement between the protagonists about the nature of appropriate evidence and its import. Indeed, *inside* Wales, there has also developed a conventional wisdom that the Welsh education system is under-performing significantly, based on much the same sorts of evidence as that adduced by critics outside of Wales. Certainly, the Welsh media have become highly adept at delivering the message that Welsh schools and their pupils lag behind their peers in other parts of the UK and internationally, and make frequent reference to evidence from PISA, in particular, to justify this analysis.

Equally – and somewhat ironically, given the frenetic nature of the political arguments – the Welsh Government has itself acknowledged the need for substantial reform in Wales, especially following what was deemed to be a dramatically poor performance in PISA 2009. The then Minister in the Welsh Government, Leighton Andrews, in a speech delivered in February 2011, outlined a twenty-point programme of educational reform that he deemed to be necessary to raise the performance of Welsh schools to acceptable international standards; indeed, he set a target (which he has subsequently come to regret) of Wales reaching the top twenty of the PISA rankings by 2015.[8]

Again somewhat ironically, the effect of this reform programme has been to bring many aspects of Welsh educational policy into much closer alignment with the policies being pursued in England. More specifically, 'school improvement' has become the cornerstone of policy in Wales, with policies aimed at raising standards of literacy and numeracy for all pupils, using individual data to track pupils'

7 PISA is the only international study of this kind in which Wales participates.

8 L. Andrews, 'Teaching Makes a Difference', speech by the Minister for Children, Education, Lifelong Learning and Skills, 2 February 2011, Cardiff; L. Andrews, *Ministering to Education*.

progress more carefully, and enabling new forms of accountability through a new inspection format and the publication of performance data for schools on an annual basis being prioritized. Similarly, enhanced leadership in all areas of the education system, the development of more effective forms of support for and collaboration between schools (to share 'good practice'), and the improvement of the quality of teachers and their teaching have come to be seen as key mechanisms for raising Welsh attainment standards. All of this mirrors rather closely much of the approach adopted in England (and more widely). However, significant differences do remain, especially with respect to the organization of (especially secondary) schooling (there are no academies or free schools and few foundation schools in Wales); the nature of the National Curriculum (most importantly in relation to provision during the early years of primary schooling and 'core' subjects at GCSE and, increasingly, GCE A level); and the system of qualifications (especially in relation to GCSEs and A levels, as well as the Welsh Baccalaureate).[9]

What this brief outline emphasizes, therefore, is that the narrative of a Welsh education system that is under-performing has had major impacts not only on the ways in which schools in Wales are *perceived* both inside and outside the country, but also – and increasingly – on the *actual* form of educational policy in Wales. Accordingly, it is important to understand the extent to which the account of an education system in Wales that is in 'crisis' is founded on robust analysis. This, in turn, entails subjecting the evidence on which it is based and the inferences drawn from it to careful scrutiny.

International Comparisons of Educational Attainment

As we have seen, the comparison of levels of educational attainment in Wales with those in other parts of the UK and internationally have provided an essential foundation for the narrative of an underperforming educational system. PISA is firmly entrenched as the most important means of measuring this relative performance of the Welsh educational system. Wales has experienced its 'PISA shock', following what were seen as the very disappointing results of PISA 2009;[10] and this has led to significant shifts in the nature of the Welsh Government's education policy. The official commitment to raising Wales' position in the PISA 'league tables' by 2015 remains in place as an underpinning for much of the reform currently being undertaken; this is despite the fact that, in the light of the results of PISA 2012, the current Welsh Minister of Education and Skills, Huw Lewis, announced recently that the target is to be framed in terms of attaining average scores of at least 500 in reading, mathematics, and science, rather than reaching the top twenty. It is also instructive that 'core' GCSEs in English, first-language Welsh, and Mathematics in Wales are currently being revised in order to align them more

9 G. Rees, 'A Crisis in Welsh Education? New Approaches in Harsh Times', *Education Review*, 24.2 (2012), 40–49.

10 Many countries have experienced equivalent 'PISA shocks' and have generally responded with significant educational reforms.

closely with what the PISA surveys test. Accordingly, to understand the dominant account of the current state of Welsh education, it is necessary to examine PISA itself.

PISA is a survey of the educational attainments of 15-year-olds undertaken by the OECD in a large number of countries across the world (some 65 countries participated in PISA 2012, with the overwhelming bulk drawn from the members of the OECD and the European Union). It has taken place every three years since 2000; and Wales has participated in its own right since 2006 (although only results for the UK as a whole are reported for many aspects of the survey). PISA assesses student skills in mathematics, reading and science, with an emphasis on the understanding of concepts through their application in real-life contexts. The explicit objective of PISA is to facilitate valid comparisons between countries in terms of their educational performance. Therefore, results are reported only at the country level (separately for the UK and the four home countries), not for schools or individual students.[11] Public attention, as well as that of politicians and other policy-makers, tends to focus on the simplest summary outcomes: the mean values of each country's students' test scores in the three domains, and the resulting ranking of these in PISA 'league tables'.

On this basis, the dominant narrative characterizes the Welsh education system in terms of two key themes. Firstly, Welsh 15-year-olds *underperform* relative to those in other countries and to the other parts of the UK, more specifically. This is reflected in mean scores for each of the three domains that are substantially below the OECD average and the scores of the other UK countries. Wales has fewer high achievers and more low achievers than the OECD average and the other UK countries. A significant proportion of Welsh 15-year-olds in 2012 did not achieve the minimum levels of mathematics and reading skills that the OECD deems necessary for them to function effectively as competent adults.[12] Secondly, the dominant narrative emphasizes that, despite the best efforts of schools and the plethora of educational reforms introduced by the Welsh Government, Wales' performance in PISA has *deteriorated* over time. Hence, mean scores in mathematics and science were lower for the 2012 survey than they had been in 2006. In reading, the mean scores remained pretty much static over this period. Correspondingly, the rankings in the PISA 'league tables' went down between 2006 and 2012 (although this reflects the entry of new countries into PISA, as well as the mean scores attained).

Table 1 shows the mean scores in each of the three domains for the four home countries between 2006 and 2012, and adds the overall figures for the UK and the OECD average. The pattern of Welsh 'under-performance' on these measures emerges very clearly. Although the differences are relatively small in numerical terms, they are – for the most part – statistically significant (an issue to which we return later).

11 From 2015, it will be possible for schools in Wales (and England) to obtain their own PISA scores.

12 R. Wheater, R. Ager, B. Burge, and J. Sizmur, *Achievement of 15-Year-Olds in Wales: PISA 2012 National Report,* OECD Programme for International Student Assessment (Slough: NFER, 2013).

Table 1: Mean Scores for Mathematics, Reading and Science, UK Countries, UK and OECD, 2006-2012

Country	Subject	2006	2009	2012
		Score	Score	Score
Wales	Reading	481	476	480
	Maths	484	472	468
	Science	505	496	491
England	Reading	496	496	500
	Maths	495	493	495
	Science	516	516	516
Northern Ireland	Reading	495	499	498
	Maths	494	492	487
	Science	508	511	507
Scotland	Reading	499	500	506
	Maths	506	499	498
	Science	515	514	513
UK	Reading	495	494	499
	Maths	495	492	494
	Science	515	514	514
OECD Average	Reading	492	494	496
	Maths	498	496	494
	Science	500	501	501

It is undoubtedly the case that PISA provides one important measure of the educational attainment of young people in Wales; certainly, the decision of the Welsh Government to enter Wales into the survey in its own right may be justified in terms of the new information that has thereby been generated.[13] However, as with any other measurement of educational attainment, PISA has significant limitations, as well as strengths. Only by acknowledging this is it possible to *interpret* the implications of PISA results adequately. It is striking that public debate in Wales has almost wholly ignored these limitations, treating PISA as if it were entirely definitive.[14] This is somewhat surprising, as, more generally, the methods adopted by PISA to measure young people's educational attainments have become increasingly subject to critical evaluation. In fact, in the run-up to the publication of the PISA 2012 results, these methodological issues even escaped from the pages of academic publications into the – relatively – popular media; for example, Professor David Spiegelhalter, of the Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge, introduced a whole programme on BBC Radio 4 devoted to exploring these matters ('PISA – Global Education Tables Tested',

13 No account is taken here of the – as far as we know, undisclosed – financial costs to the Welsh Government of participation in PISA. Clearly, there are important opportunity costs involved here. (See http://www.oecd.org/pisa/aboutpisa/howtojoinpisa.htm for some indication of financial costs.)

14 The uncritical reception of PISA results in Wales matches that in many other countries too: see, for example, S. Grek, 'Governing by Numbers: the PISA "Effect" in Europe', *Journal of Education Policy*, 24.1 (2009), 23–37.

Square Dog Radio Production for BBC Radio 4, 20 November 2013).

To begin to appreciate the significance of these issues, it is necessary to explore some of the slightly more technical aspects of how the PISA surveys are conducted. Hence, it is probably not widely appreciated that each PISA survey is based on a *sample* of secondary schools in each country (including Wales), selected to be representative of the total population of schools.[15] Schools are able to refuse the invitation to participate; and may then be replaced by other schools from back-up samples, ensuring that the final sample conforms to the criteria set by the PISA managers. Within each school, a further sample of 30 students is randomly selected to complete the tests; and again a set proportion of those selected are required to undertake the assessments for the school's results to be included in the overall analysis.

Because participation in the PISA tests is voluntary, the nature of response and non-response is crucial. In general terms, the key issue is whether non-respondents (those schools and/or individuals who are selected, but do not participate) are atypical, as this would introduce systematic bias into the results. Unfortunately, this is not an issue that is explored in any depth in the materials published by the OECD; and hence it is not feasible to explore the possible effects here.

The OECD does provide information on the school response rates in terms of its sample of selected schools (with and without substitution from the back-up samples). In Wales, the minimum response rates for schools have been achieved (indeed, substantially exceeded) for each of the three surveys that have been conducted since 2006.[16] This contrasts with England, where these response rates have been much lower (especially in 2000 and 2003, and to a lesser extent in 2009). Indeed, some commentators have suggested that the latter problems resulted in the over-estimation of the test scores in England for 2000 and 2003 by between four and 15 points.[17]

Even if response rates are wholly unproblematic, it is clear that PISA's sampling procedures imply that only a *small minority* of the 15-year-olds in any country are selected to participate in the tests. There is, of course, no problem in principle with adopting what is a perfectly standard procedure for conducting a survey, although this is probably at variance with most people's understanding of how PISA testing is carried out. However, it does mean that the sampling itself becomes a matter of great interest.

Simply because they are derived from a sample, the results have error ('sampling error') associated with them. Hence, the mean scores for each country – and the rankings in the 'league tables' based upon them – are, in reality, estimates within a *possible range* of scores. This issue of systematic error in PISA results is further

15 The sample is also stratified according to a number of criteria. For PISA 2012 in Wales these included: maintained/independent school; region of the country; gender of the school pupils; level of GCSE performance; and local authority.

16 Pupil response rates are reported only for England, Northern Ireland and Wales combined.

17 J. Micklewright, S. Schnepf, and C. Skinner, 'Non-response Biases in Surveys of School Children: The Case of English PISA Samples', *DoQSS Working Paper 10-04* (London: Institute of Education, 2010). However, some caution is required here, as in 2000 and 2003 the results for 'England' actually included a proportion of Welsh young people.

compounded by the fact that not all of the young people who participate in the PISA surveys actually answer all the test questions. The issue here is that each survey focuses on *one* of the three domains: in 2006, it was science; in 2009, it was reading; and in 2012, it was mathematics. In 2015, it will be science again.[18] Only in the principal domain are the respondents required to answer all questions; in the two 'minor' domains, respondents are asked to attempt only a proportion of the total number of questions comprising the tests. Indeed, Kreiner and Christensen calculate that in the 2006 survey, almost half of the respondents did not answer any reading questions at all; and only some 10 per cent were tested on all items.[19] This introduces a further element of error, making the range of possible scores around the reported mean larger than it would otherwise be.

The OECD acknowledges this (albeit not very prominently), but it gets lost in the public reception of PISA results. For example, in the Appendices of the official country report for Wales for PISA 2012, the OECD's estimates of the error and of the associated range of mean scores are set out. Here, it is shown that, whilst the mean score for mathematics, for instance, is reported to be 468, it is more accurate to report that were the sampling to be repeated on numerous occasions, the 'true value' will fall within the interval (between 465.8 and 470.2) 95 per cent of times. Similarly, we can be confident that the 'true value' for reading will lie between 477.3 and 482.7 for 95 per cent of times (the reported mean score was 480); and for science between 488.0 and 494.0 (with a reported mean score of 491).[20] Although these ranges are quite small, given that the numerical differences between reported means are also small, treating the scores in this more accurate way can have substantial implications for rankings in the PISA 'league tables'. Moreover, the errors and consequent ranges of possible scores are substantially larger for many countries other than Wales.

What is especially controversial, however, is the way in which the results for the 'minor' domains – in which, as we have seen, a substantial proportion of respondents do not answer questions – are treated. This problem is addressed by the use of an established statistical technique (the simple Rasch model) to generate multiple 'plausible values' for each respondent, which are estimates of what each respondent would have scored if they had actually completed the tests, based on their answers to other questions (which they have completed) and a set of known characteristics of the respondent. These 'plausible values' are subsequently incorporated into the calculation of the mean scores reported for countries, as if they were actual test responses.

Again, there is no problem in principle about using imputed ('plausible') data in this way. However, this procedure does require that the specific application of the Rasch model to PISA is appropriate. In a recent paper in *Psychometrika*, Kreiner and Christensen address this issue. They argue that the presence of

18 Only in 2015 will the first 'cycle' of surveys in Wales be completed, providing a somewhat more robust test of change over time.

19 S. Kreiner and K. Christensen, 'Analysing Model Fit and Robustness: A New Look at the PISA Scaling Model Underlying the Ranking of Countries According to Reading Literacy', *Psychometrika*, 79.2 (2014), 210–231.

20 Wheater *et al., Achievement of 15-Year-Olds.*

substantial Differential Item Functioning (DIF) in the PISA data, resulting from the fact that questions have different difficulty in different countries, violates some of the conditions necessary for the application of the simple Rasch model to be appropriate. Hence, the 'plausible values' on which a substantial part of the PISA results is based do not provide valid estimates of what the young person would have answered if they had actually responded to all the questions. They conclude that it is not possible to estimate what the effects of using a substantial amount of imputed data of this kind, derived from an inadequate model, are on the subsequent detailed analyses that produce the mean scores and country rankings that are reported by OECD.[21]

What conclusions, therefore, should be drawn from this rather technical discussion? The criticisms of the appropriateness of applying the simple Rasch model to impute PISA data are fundamental. If these criticisms are robust, they quite simply undermine the approach adopted in PISA and the results that are produced. At the time of writing, the weight of expert judgement clearly supports this critical evaluation of the PISA analysis. Accordingly, there must remain major doubts about the validity of the country mean scores and rankings that are derived from PISA and which currently provide a major basis of policy-making in many countries, including Wales.

It can be argued, of course, that the most reliable cross-national comparisons on the basis of PISA results are those between the home countries of the UK. Here, it might be suggested that the confounding effects of differences in language, as well as in educational and wider cultural context, on the comparability of test responses are minimized. In this context, therefore, it is certainly notable that, in general, 15-year-olds in Wales have performed significantly (in a statistical sense) below the levels achieved by their peers in the other UK countries. In PISA 2012, for example, the Wales scores for mathematics, reading, and science were all significantly lower than those for all the other UK countries, whilst the differences between the scores of the other UK countries were not significant, with the single exception of Scotland's higher score in mathematics compared with Northern Ireland.[22]

However, the interpretation of the *implications* of these findings is much more problematic than is frequently acknowledged. In fact, the PISA surveys are able to provide only limited guidance here. PISA's strength is in providing a *snapshot* of the educational attainments of 15-year-olds; it is a *cross-sectional* survey. We need to be able to track individuals through the educational system in order to be able to sort out the effects on educational attainment of different types of factor, such as the characteristics of the educational system and the wider social and economic environment in which individuals grow up. As for PISA, its fallibilities and limitations suggest its outcomes should be seen as no more than a starting point for further analysis, rather than a definitive account of the state of health of a country's educational system. Certainly, 'moral panics' based on PISA scores can be seen to be unwarranted and, ultimately, counter-productive.

21 Kreiner and Christensen, 'Analysing Model Fit'.
22 Wheater *et al.*, *Achievement of 15-Year-Olds.*

'Home International' Comparisons of Educational Attainment

The second major sort of evidence that underpins the narrative of an underperforming education system in Wales comprises the measures of attainment that are generated as part of the 'normal' activity of schools. Most important here are the assessments that are used to track the progress of individual pupils through primary and secondary schooling. These assessments are carried out at the end of each Key Stage, culminating in GCSEs at the completion of Key Stage 4, the final part of compulsory education. It can be argued that assessment data of this kind are especially powerful, in that schools, teachers and pupils undoubtedly take them seriously. Evaluations of school performance are often based upon levels of attainment achieved in these terms;[23] teachers are judged at least partly by their pupils' 'results'; and pupils' own access to further and higher education or employment is dependent upon their individual performance and the qualifications they gain. In short, the stakes are high in relation to these assessments, in a way that has not always been clear about PISA.

In the debates about the relative underperformance of the Welsh education system, it has been attainment measured at the end of compulsory education in terms of GCSEs (and equivalent qualifications) that has been most prominent. On this basis, therefore, it has been possible to compare attainment in Wales with that in England and Northern Ireland (although comparisons with the latter are only very infrequently made). Scotland's attainment levels are much more difficult to incorporate, in that they are measured in terms of the Scottish qualifications framework, which is significantly different from what is operated in other parts of the UK.

If one compares Wales with England in these terms, it is clear that, from the early 2000s, there was a progressive widening of the shortfall between Wales and England in terms of the standard measure of attainment at the end of compulsory schooling, known as the Level Two Threshold. Figure 1 plots this increasing differential in the proportion of young people attaining the Level Two Threshold. As can be seen, by the end of the decade, the gap between Wales and England in these terms had widened to some 10 percentage points, although it was significantly smaller when a narrower definition of the Threshold, including Mathematics and English, is used.

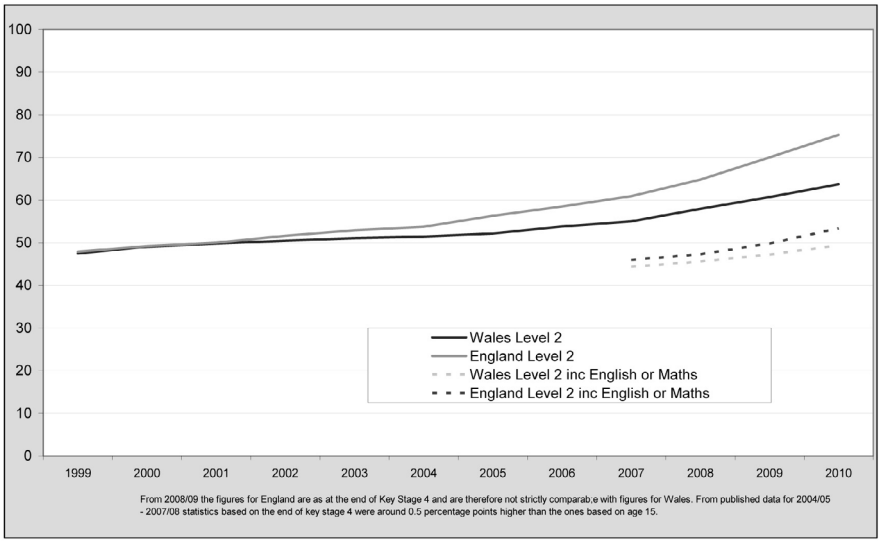
Evidence of this sort has provided a basis for politicians and other policy-makers, as well as commentators more generally, to portray Welsh educational performance in negative terms.[24] However, it has also been used in more detailed research

23 This is true of both the school inspections carried out by Estyn in Wales and the annual 'banding' of secondary schools by the Welsh Government. Data of this kind will remain an essential element of the new system of 'banding' recently announced: Welsh Government, *Qualified for Life: An Education Improvement Plan for 3 to 19-year-olds in Wales* (Cardiff: Welsh Government, 2014).

24 J. Blair, 'Devolution and Divergence in Secondary Education in Wales: Measuring Performance' (unpublished MSc (Econ) dissertation, Cardiff University, Cardiff, 2012).

analyses to illuminate the character of Welsh educational underperformance.[25] In this context, an especially influential study has been that by Burgess, Wilson and Worth, which argues – on the basis of sophisticated statistical analysis – that the widening gap between Wales and England is attributable to the decision of the Welsh Government in 2001 to abolish the publication of 'league tables' of school attainment levels, thereby removing an important source of information for parents and reducing the pressure on schools in Wales to perform to their best ability.[26] Interestingly, politicians – inside and outside Wales – who are opposed to the approach adopted by the Welsh Government, as well as other policy-makers and commentators, have subsequently taken up this study.

Figure 1: % Attaining the Level 2 Threshold, Wales Compared with England, 1999-2010



Source: Graph provided by the Welsh Government.

In a robust critique of the study, Goldstein argues that it is just as plausible that it is the lack of experience in taking tests during earlier Key Stages (which were also abolished in Wales, along with the 'league tables') that explains the gap between the examination performance of Welsh and English pupils.[27] However, more pertinent to our concerns here is his contention that because pupils in Wales

25 D. Reynolds, 'New Labour, Education and Wales: The Devolution Decade', *Oxford Review of Education*, 34.6 (2008), 753–765.

26 S. Burgess, D. Wilson, and J. Worth, 'A Natural Experiment in School Improvement: The Impact of School Performance Information on Pupil Progress', *Journal of Public Economics*, 106 (2013), 57–67.

27 H. Goldstein, 'Do league tables really improve test scores?' (2014) http://hgeduc.blogspot.co.uk, accessed 16 September 2014. This factor may also affect Welsh performance in PISA assessments.

are predominantly assessed for their GCSEs by different examining boards from pupils in England, Burgess and his colleagues are not comparing 'like with like'.

Much more significantly, however, it is necessary to understand the nature of the qualifications that are being compared. The Level Two Threshold is frequently described as the achievement of five A* to C grade GCSEs. In fact, its full definition is five A* to C grade GCSEs or *equivalent* qualifications. That is to say, measures of pupil attainment of this kind include not only GCSEs but also a wide range of vocational qualifications, including BTECs and many others. This definitional detail is significant because if we separate out GCSEs from vocational qualifications, a rather different picture emerges from that which provides the basis for Burgess *et al.*'s analysis (as well as numerous other, far less measured accounts of the gap in performance between Wales and England).

Figure 2 illustrates this. When only those pupils attaining five A* to C grade GCSEs alone is compared, it can be seen that the Welsh performance is almost indistinguishable from that in England. The difference in the total percentages achieving the Level Two Threshold is accounted for by the much higher numbers in England achieving this level by virtue of vocational qualifications. It is also striking that, over the past couple of years, a small gap has opened up between the Welsh and English proportions of young people attaining the Level Two Threshold through GCSEs alone.[28] Simultaneously, however, the difference in the overall figures has been narrowing significantly, as more young people achieve the Threshold on the basis of vocational qualifications.[29]

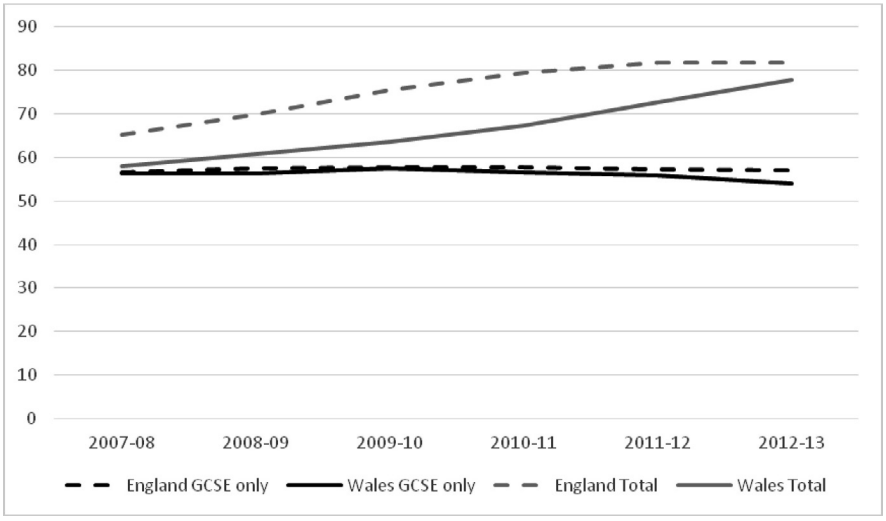
It is difficult to reconcile these trends with the conclusions drawn by Burgess and his colleagues. Clearly, one would expect the impact of the removal of 'league tables' to be felt on GCSE scores, as well as overall levels of attaining the Level Two Threshold. It would appear that, rather than the availability of information to parents, it is the pattern of entry to different types of qualification that is crucial. During the early part of the period, pupils in Wales were simply not attaining vocational qualifications to anything like the same extent as their English peers; and this accounts for the overall gap in performance, measured by the proportion of young people attaining the Level Two Threshold. More recently (largely after the period considered by Burgess *et al*.[30]), there has been a significant increase in the percentage of Welsh pupils achieving the Threshold on the basis of vocational qualifications and a simultaneous decrease in the equivalent English proportions (presumably reflecting the changes in the UK Government's approach in England,

28 This is larger when the measure includes English, Welsh First Language, and Mathematics.

29 It should be noted, however, that after 2009–10, the Welsh Government no longer reports these comparisons in its statistical outputs, on the grounds that the basis of calculation of the proportions attaining the Level Two Threshold in the two countries is too divergent. A comparison is provided, however, in Programme for Government reports, although this will not be possible after 2014 given the changes to vocational equivalences, discounting rules, and GCSE grading coming into effect in England.

30 Also during the period after that considered by Burgess and his colleagues, the Welsh Government has established the award-winning web-site, mylocalschoolwales.gov.uk, which provides an extensive range of performance information for every state school in Wales.

Figure 2: % Achieving the Level 2 Threshold, Total and by GCSE only, 2007-08 to 2013-13



Source: from data supplied by the Welsh Government

following the recommendations of the Wolf Report[31] and the introduction of the English Baccalaureate). Ironically, this increase in the overall proportion of Welsh young people achieving the Level Two Threshold has been paralleled by a small decrease in those attaining this level by GCSEs alone.

Certainly, there is no straightforward support here for the narrative of significant under-performance in the Welsh education system (relative to that of England). Rather, interesting and important *questions* are raised. For example, it is difficult to interpret the implications of the shifts in the proportions of young people in Wales attaining the Level Two Threshold on the basis of vocational qualifications without knowing a great deal more about the nature of the vocational qualifications and patterns of access to them. More generally, the discussion underlines the extent to which the results of 'home international' comparisons of this kind are dependent on the *exact* nature of the measures adopted. More work is required to resolve the complicated issues of comparability between Wales and England (and the other UK countries) in respect of the measures on which such comparisons are made. Even where the same terms are used (Level Two Threshold, GCSEs, etc.), it remains to be established precisely how these are applied in Wales and in England (and Northern Ireland). The contrasting changes in the content of the GCSE curriculum currently underway in both countries will make establishing a common basis of comparison even more difficult in the future than it is currently. Moreover, the effects of these curriculum changes will be compounded by the progressive adoption of a radically different grading system for GCSEs in England from 2017.

31 A. Wolf, *Review of Vocational Education: The Wolf Report* (London: Department for Education and the Department for Business, Innovation and Skills, 2011).

Even if we set these measurement issues to one side, there remain difficult issues relating to the *interpretation* of the differences in levels of educational attainment uncovered by these 'home international' comparisons. In particular, the observed differences in attainment levels may be attributable to a wide range of economic and social factors, other than the quality of the educational system itself. Accordingly, the key here is that, as far as possible, we should isolate the impacts of the education system from those of the wider economic and social environment. Clearly, simple comparisons effectively ignore the fact that there are substantial differences in economic and social conditions between the constituent countries of the UK and that it is very likely that these influence observed levels of educational attainment significantly.

A number of approaches have been adopted to address this issue. For example, the Welsh Government has analysed key aspects of the performance of the Welsh education system by focusing on contiguous local authority areas on either side of the Wales-England border, thereby attempting to compare 'like with like'. An alternative method has been to undertake regional analysis. The Welsh Government has carried out a number of analyses that compare educational performance in Wales with that in the standard regions of England, rather than simply with England as a whole. Clearly, approaches of this kind are very limited in the extent to which they enable 'like with like' comparison.

An alternative method, which – as far as we know – has not been adopted previously in 'home international' comparisons, is to use a statistical technique known as propensity score matching (PSM). We suggest that PSM retains the strengths of more conventional statistical approaches (based on traditional forms of regression analysis) in allowing us to separate out the effects of different kinds of factor on educational outcomes, but also has significant advantages over these more conventional methods, especially in relation to the clarity with which results can be presented publicly, to policy-makers and within civil society more widely.

What follows is intended to illustrate the *potential* of this methodological approach. We use PSM to match the 22 local authority areas in Wales with the 22 local authority areas in England that most closely correspond to them, based on a set of shared economic and social characteristics. This sub-set of English local authorities may thus be thought of as a 'synthetic Wales': the closest English equivalents to the Welsh local authorities. For these illustrative purposes, this matching exercise is based on a small number of shared characteristics: age structure (% retired); ethnicity (% White British); population density (people per hectare); and socio-economic group composition (% in different NSSEC groups).[32] Clearly, these could be added to in order to produce a closer alignment of the comparator English local authorities with the actual Welsh ones.

It is also important to note that this PSM comparison operates at an *aggregate* level, using local authority areas as its units of analysis. This risks understating the heterogeneity of the individuals within each area, thereby obscuring the true nature of the individual (pupil) level relationships. In strictly statistical terms,

32 We are very grateful to our colleagues at WISERD, Rhys Davies and Sam Jones, for their significant contributions to this analysis.

therefore, it can be argued that it would be better to base the analysis on *individuals*, thereby avoiding the 'ecological fallacy'. We would argue, however, that there are significant *presentational* advantages in adopting the notion of a 'synthetic Wales'. In addition, we have used a measure of attainment based on GCSE passes alone. Whilst this addresses some of the difficulties associated with producing comparable attainment data that were noted earlier, further research would also be required to produce a more robust set of comparable measures, especially in light of the increasing divergences in curriculum and assessment systems between the countries of the UK that are currently in train.

Table 2 presents the results of the PSM analysis. As can be seen, the effect of this most basic attempt at matching the Welsh local authorities with an *equivalent* set of local authorities in England is to reduce appreciably the gap in attainment, from 8.6 percentage points (that is, between Wales and England) to 5.7 percentage points (between Wales and 22 Equivalent English Local Authorities). Accordingly, even this initial attempt at ensuring that we are comparing 'like with like' (and thereby 'holding constant' the effects of differences in economic and social conditions) indicates that understanding the nature of the relative performance of the Welsh educational system is more complex than is suggested by a simple narrative of significant underperformance in relation to that in England (and the other countries of the UK). Of course, even on the basis of the matched comparison, there remains a substantial shortfall in Welsh performance. It is possible that a more sophisticated approach to matching would reduce this gap still further. However, further research would be required to explore this fully.

Table 2: % Pupils at the End of Key Stage 4 Achieving Level Two Threshold (including English/First-language Welsh and Mathematics) on the basis of GCSEs alone, Wales, England and 22 Equivalent English Local Authorities, 2012-2013

Wales	England	Equivalent English Local Authorities
53.0	61.6	58.7

Source: calculated from the National Pupil Database.

Concluding Comments

Our aim in this paper has been to raise issues by evaluating the robustness of evidence, rather than to present the fully developed conclusions of completed research. We have shown that what has become the established narrative of the significant underperformance of the Welsh education system (and schools, in particular) is founded on a far less secure evidence base than is generally acknowledged. For a variety of mostly technical reasons, the evidence drawn from cross-national comparisons (both between the countries of the UK and internationally) is much less robust and more indeterminate than is acknowledged, not only by politicians and other policy-makers, but also in the public debates that their pronouncements – and their presentation in the media – provoke. Indeed, even much more sober,

research-based analyses frequently ignore the complexities in the evidence that we have outlined.

We hope that students, parents and educational professionals may draw some comfort from our more rigorous evaluation of the evidence about educational attainment in Wales. However, we should emphasize here that we are certainly *not* suggesting that everything in the Welsh educational garden is rosy and that there are no major concerns about the performance of the educational system in Wales. Despite the substantial reservations that we have outlined about the nature of the evidence, it remains the case that there is at least a *prima facie* case that attainment levels in Wales lag behind those in comparator countries to some degree, although the extent of this may well be overstated in the dominant narrative.

However, whilst public concern about Wales' educational system is entirely legitimate, there is currently a danger that simplistic readings of PISA and other external benchmarking of Welsh educational performance are serving to close off debate, rather than to open up new avenues of educational development. Sustaining the latter requires a more open and enquiring approach to the nature of the evidence and the sorts of analysis that it makes possible than is promoted in Wales (and more widely) at present. Clearly, the possibilities for such open and sceptical analysis are very extensive. However, such analysis will be very difficult to sustain unless systematic attention is paid to the development of a more robust evidence base.

We can identify two areas here that are in need of urgent further work. Firstly, there is a need to develop a more effective database that will permit the *longitudinal* analysis of educational attainment in Wales. This is necessary if we are to be able to understand the real impacts of the educational system itself, whilst taking account of the wider social and economic conditions that also influence educational outcomes. One possibility here would be to make much more systematic use of the Millennium Cohort Study, especially where this is combined with administrative data on individuals derived from the National Pupil Database. It is to be hoped that, in due course, the latter will be enriched by the incorporation of results from the newly instituted testing of literacy and numeracy.

Secondly, as the qualifications systems of the constituent countries of the UK draw further apart (in terms both of curriculum and assessment), a much more intensive effort will be required to establish the comparability of attainment measures that are based on these qualifications. We have addressed some of the difficulties that arise in using comparative data when analysing *past* patterns of educational attainment. However, these will be severely compounded in the future. Accordingly, if external benchmarking of the Welsh educational system is to be retained as a principal mode of evaluation, then establishing the comparability of what are becoming different qualifications in the home countries of the UK is essential. Whilst this will require a significant research effort, abandoning 'home international' comparison (which is what appears to be happening currently) is not an option, given that the major alternative of international benchmarking through PISA is beset by major technical difficulties (quite apart from the fact that it takes place only every three years).

More generally, it is important to recognise the limitations of an almost exclusive dependence on the *external* benchmarking of Wales' educational system

as a guide to the development of policy and the improvement of educational provision. For example, even if we accept PISA results at face value, it is not clear what policy implications follow from them. Whilst the OECD itself continues to draw general conclusions about the most effective way to organize educational provision on the basis of PISA results, the extent to which these conclusions are genuinely warranted by the PISA evidence remains a matter of debate. Further research would be required to explore *fully* the implications of key features of the Welsh education system, such as, for instance, the lower levels of expenditure per pupil in Welsh schools.

Certainly, the policy 'lessons' that are drawn here remain at a very high level of generality.[33] In part, at least, this reflects the fact that the high-scoring countries are very heterogeneous in the nature of their educational provision. Moreover, many of them have social environments and, indeed, political regimes that could not be transferred to Wales. Policy-makers in Wales certainly need to be outward-looking in terms of their deliberations as to what is required to make the Welsh education system more effective. However, it is also important not to lose sight of the fact that devolution provides the constitutional space to set educational priorities that are specific to Welsh circumstances.

In this context, policy-makers would do well to consider that the relationships between improved levels of qualifications and/or the cognitive skills measured in PISA and other measures of educational attainment, on the one hand, and enhanced economic growth, on the other, are in reality highly complex. Increasing 'human capital' will lead to improved economic performance only in the long run and, crucially, only if businesses and other organizations adapt themselves to make full use of this resource. Improving productivity and rates of innovation are key to a competitive economy; increasing educational attainment levels may be necessary, but is not sufficient to ensure that these are achieved. Indeed, perhaps we all need to remind ourselves that high-quality education is about much more than the sorts of educational attainment that currently drive policy development in Wales (and elsewhere).

33 For example, OECD, *Improving Schools in Wales: An OECD Perspective* (Paris: OECD, 2014); A. Schleicher, 'Qualified for Life: Embedding PISA Skills in Welsh Education', presentation at National Education Conference, Cardiff, 2014.