

From Local to Global: linking up the assessment and improvement agendas in Education

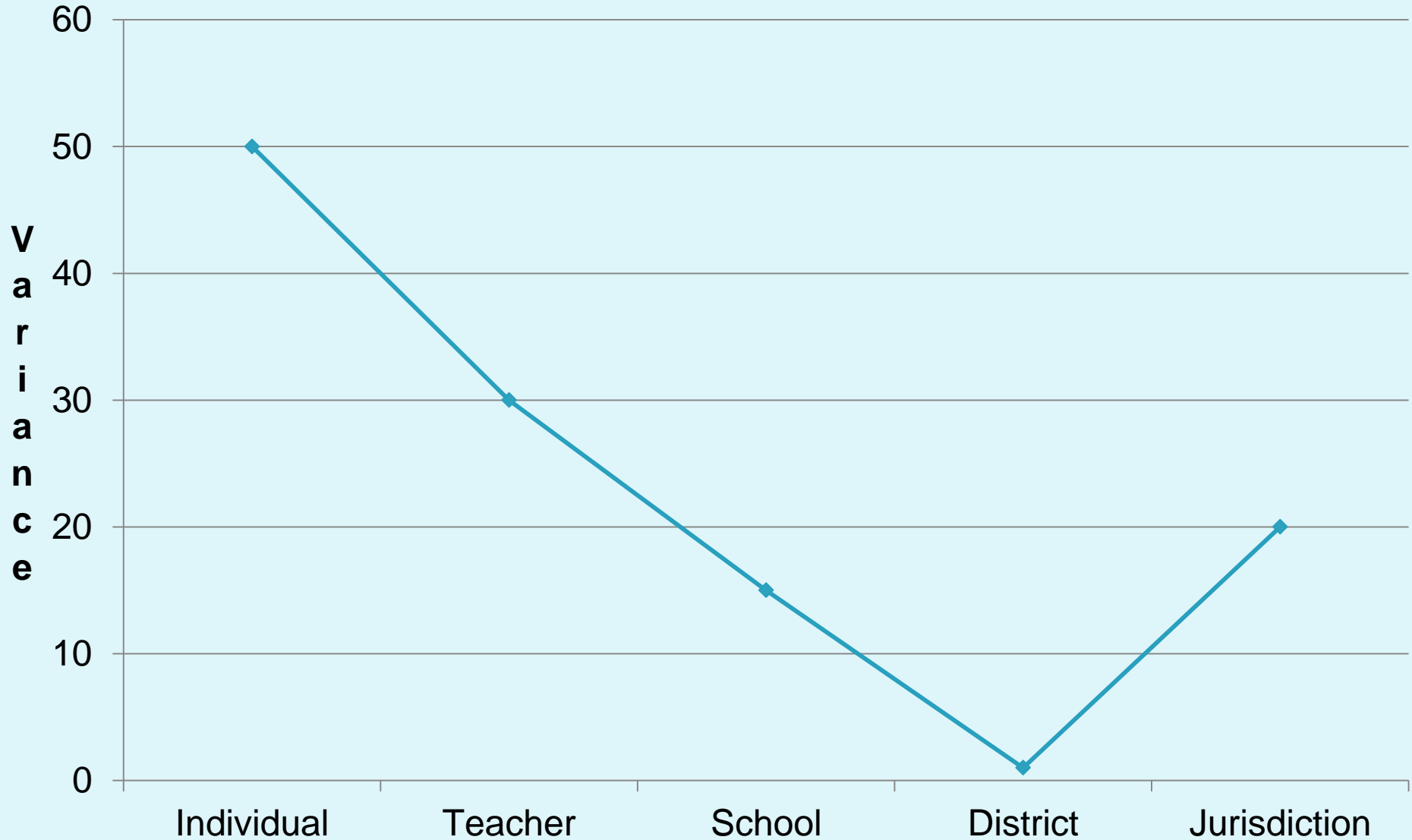
Professor David Hawker
College of Teachers and Durham University, UK

What have we learnt about
assessment and school
improvement in the past 20
years?

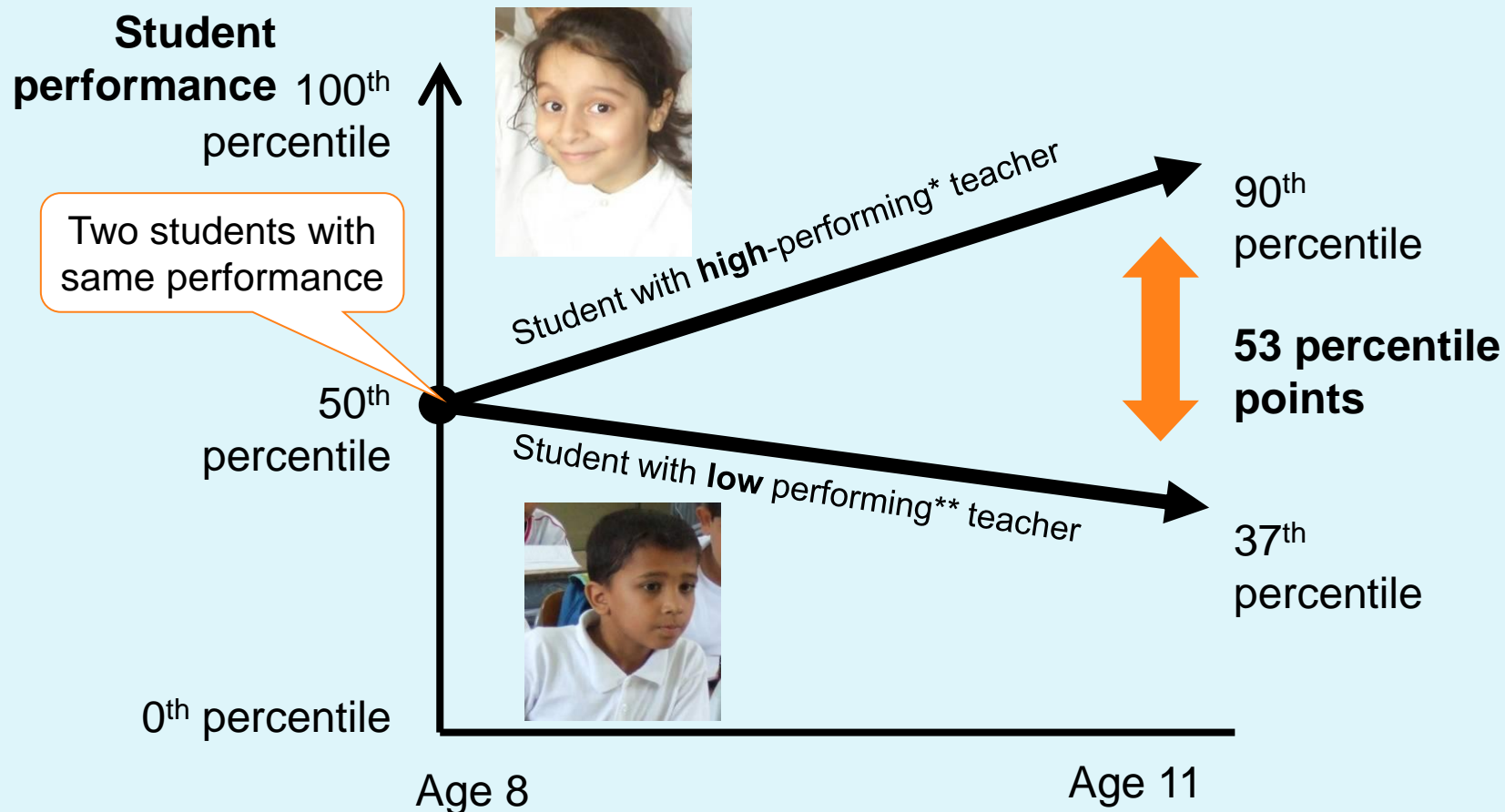
The Literature

- A student's progress is tied to his/her starting point
 - *Prior achievement is associated with 50% of the variance*
- Teachers and classes are key
 - *Up to 40% of the variance*
- Schools are important
 - *10-30% of the variance*
- Districts are of little importance
 - *1% or less of the variance*
- Educational systems (aka jurisdictions) are important
 - *Up to 20% of the variance*

Graphically



Teacher quality is the most important lever for improving student outcomes



*Among the top 20% of teachers; **Among the bottom 20% of teachers

Analysis of test data from Tennessee showed that teacher quality effected student performance more than any other variable; on average, two students with average performance (50th percentile) would diverge by more than 50 percentile points over a three year period depending on the teacher they were assigned

Source: Sanders & Rivers Cumulative and Residual Effects on Future Student Academic Achievement, McKinsey analysis

What is the research evidence about the effectiveness of different interventions?

The Education Endowment Fund in the UK has worked with Durham University to create a 'toolkit' allowing schools to evaluate different types of intervention, based on cost and impact

The data is taken from a range of studies in different countries, and an average effect size is calculated for each type of intervention, to produce a 'score' for impact

The resulting league table makes interesting reading....

The EEF toolkit league table of interventions – selected items

Intervention	cost	evidence	impact
Feedback to pupils	low	good	+8 months
Meta-cognition and self regulation	low	very good	+8 months
Peer tutoring	low	very good	+6 months
Early years intervention	very high	very good	+6 months
Small group tuition	high	moderate	+4 months
Digital technology	Very high	Very good	+4 months
Reducing class size	Extremely high	Good	+3 months
After school programmes	Very high	moderate	+2 months
Homework (primary)	Very low	good	+1 month
Teaching Assistants	Very high	moderate	0 months
Performance pay	low	weak	0 months
Selection/tracking	Very low	good	-1 month
Repeating a year	Very high	Very good	-4 months

So 'feedback' is top of the table?

Yes, and this is supported by hundreds of studies from across the world, eg

- Black and Wiliam *Inside the Black Box* 1998. Using 250 sources from around the world, the study found that giving pupils formative feedback rather than grades resulted in effect sizes of between 0.4 and 0.7 in terms of improvement in performance
- Hattie and Timperley *The Power of Feedback* 2007. Reported on 12 meta-analyses of feedback in classrooms. Average Effect Size = 0.79 (varies according to the type of feedback, eg use of cues 1.1, corrective feedback 0.37).

Hence Governments everywhere have been adopting policies on formative assessment and interactive pedagogy, not least Singapore

Good teachers are skilled in both formative and summative assessment

- They understand formative assessment as Process – an ongoing conversation between the teacher and the learner
- They understand summative assessment as Measurement – producing data which can provide high quality, sharply focussed information for evaluating the quality of outcomes

Building Assessment Literacy

If assessment is such an important driver for school improvement, it's important to ensure that all teachers and principals are well-versed in it:

- Technical understanding of assessment methodologies
- Practical classroom assessment skills
- Skill in interpreting data
- Understanding of children's learning, and how to use assessment to evaluate different pedagogical strategies

How educational assessment skills are becoming more widespread

- Professional development opportunities (eg this conference!)
- Associations of professionals, eg Chartered Institute of Educational Assessors in UK
- Formal incorporation of assessment into pre-service and in-service training programmes, eg Armenia
- Growing number of Education Masters qualifications focussing on assessment (eg NIE course in Singapore)
- Growing public debate concerning school standards, and greater sophistication in interpreting the data
- More explicit linking of assessment with pedagogy at school, with use of toolkits of benchmarked effective practice (eg OECD, McKinsey, Education Endowment Foundation)

Trends in national assessment systems

- Refinement of systems in response to perverse incentives and unintended consequences
- Growth of formative assessment practices (assessment for learning) to improve children's learning
- Increased use of assessment data in school improvement

Using assessment for school improvement

- to measure the impact of different strategies, to improve teaching and instruction
- to evaluate the success of different groups of students, to target interventions more effectively
- to evaluate performance and set targets, as part of a regime of monitoring and inspection
- as a passport (or hurdle) to the next stage in education – thus spurring schools to achieve the best results possible

Goodhart's Law (1975)

An indicator ceases to have value when it is used as a target

What does this mean?

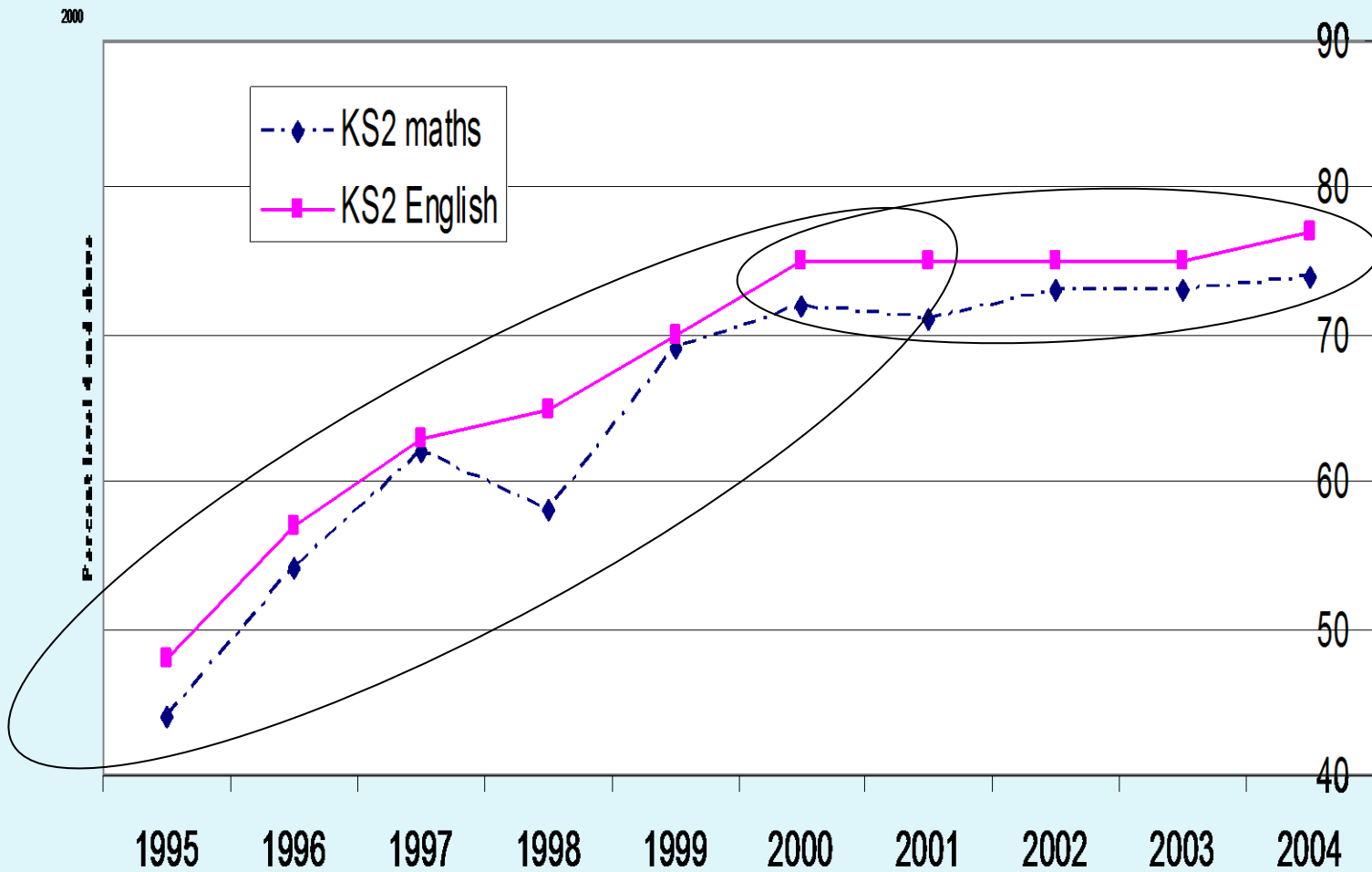
It means you can potentially use the same assessment for formative/diagnostic purposes and for national sampling of performance, but if you also try to use it as an accountability instrument at school or individual teacher level, it will inevitably become distorted.

**What's been happening in
England?**

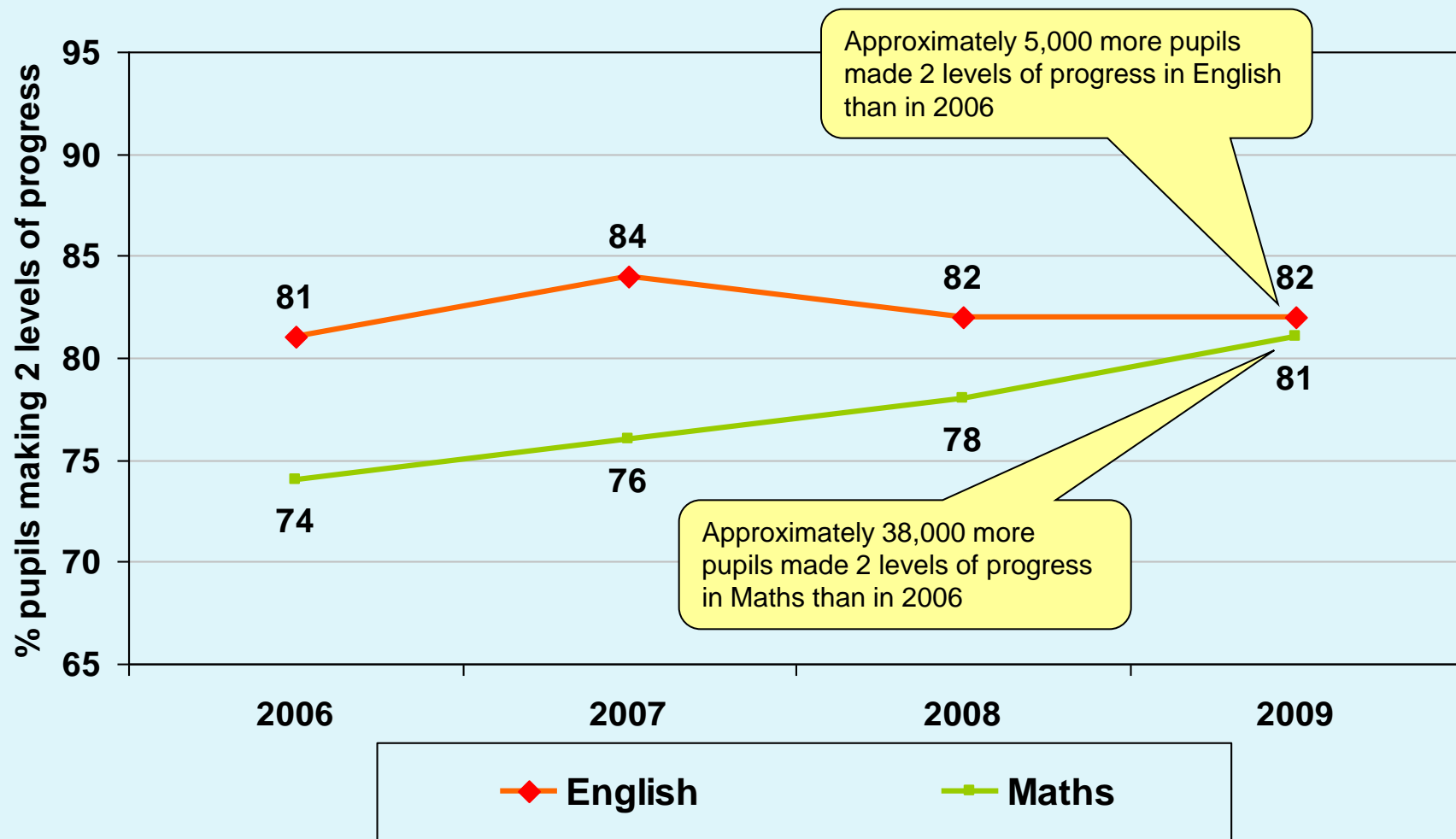
Massive efforts to raise standards

- National Curriculum
- National testing
- Ofsted
- More than 600 initiatives for Basic Skills in primary schools
- National Numeracy Strategy
- National Literacy Strategy
- League tables, target setting, homework clubs, etc etc etc

KS2 Percent With Level 4+



Change in numbers of pupils making expected progress between KS1-2 from 2006-2009



What was wrong with levels?

- Too broad for short term measurement of progress – schools needed year by year targets
- Too vaguely defined – level descriptions not precise enough (original statements of attainment discontinued)
- Meant different things in different curriculum areas – didn't work with less linear subjects
- Differently interpreted in primary and secondary sectors

Independent review of Testing and Assessment 2011

Four key principles:

1. Ongoing assessment is a crucial part of effective teaching, but should be left to schools, with no government prescription
2. External school level accountability is important but must be fair – measures of progress as well as measures of attainment
3. Wide range of school performance information should be published, to help parents and others hold schools to account in a fair and rounded way
4. Both summative teacher assessment and testing are important and should both be published

UK government 2013 proposals for Primary schools: (1) Assessment

- No levels – expectations based purely on programmes of study for each key stage
- Formative assessment entirely the school's responsibility
- Slimmed down national end of key stage tests in reading and maths – national sampling in science
- 'Secondary readiness' the key criterion
- Results expressed as standardised scores (80-130), with 100 representing 'secondary readiness', and attainment in relation to the national cohort expressed as deciles
- Progress reported against a previous baseline (either age 5 or 7)
- Summative school based assessment to be used to report children's progress annually against the new national curriculum programmes of study, but no levels or sub-levels, and no national tests

UK government 2013 proposals for Primary schools: (2) Accountability

- End of key stage tests reported both as annual results and as three year rolling averages
- Reporting of average scaled score, % of pupils matching the 'secondary readiness' standard, distribution of pupil scores across national deciles, average rate of pupil progress (value added)
- 'floor target' – 85% of pupils to reach the new 'secondary ready' standard, and/or score of 98.5-99 on value added indicator
- Additional reporting of % of pupils in top decile
- Additional reporting of progress for 'pupil premium' students

How will this help school improvement?

- More direct links to curriculum goals
- Formative assessment set free from national prescription
- Use of numerical scores to differentiate performance and raising of expectations ('secondary readiness' will be more demanding than current level 4)
- Continued use of school level 'floor targets', but with added incorporation of value added measure
- More frequent re-inspection of schools below the floor target

What are the risks?

- Narrower tests could narrow the teaching further
- Arbitrary 'secondary readiness' standard not rounded enough, nor based on empirical evidence
- Schools will adopt different approaches to assessment and reporting, making benchmarking more difficult
- Too much trust placed on the reliability of tests, and lack of insight by inspectors
- Danger of game-playing by schools

What's been happening at the international level?

International assessments

- TIMSS – maths and science, grades 4 and 8 (every 4 years since 1995)
- PISA – reading, maths, science, age 15 (every 3 years since 2000)
- PIRLS – reading and language, grade 4 (every 5 years since 2001)

The power and potential of 'big data': *'Big data is the foundation on which education can reinvent its business model and build the coalition of governments, businesses, and social entrepreneurs that can bring together the evidence, innovation and resources to make lifelong learning a reality for all'*. Andreas Schleicher, July 2013

PISA design principles

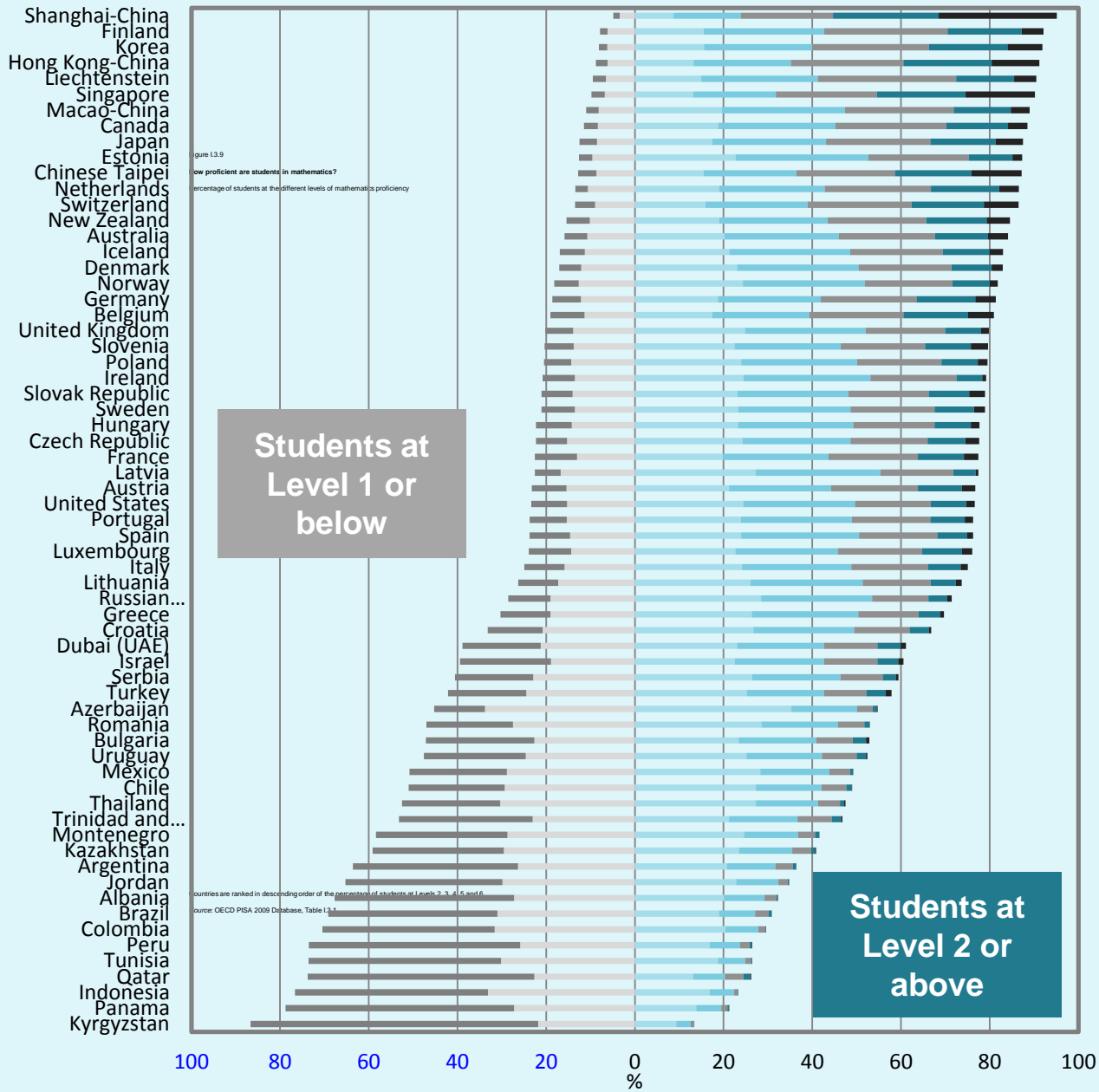
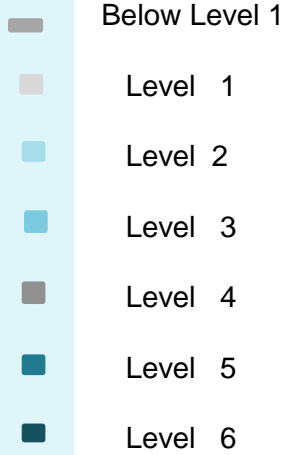
- *Public policy issues:* helping to answer questions such as "Are our schools adequately preparing young people for the challenges of adult life?", "Are some kinds of teaching and schools more effective than others?" and "Can schools contribute to improving the futures of students from immigrant or disadvantaged backgrounds?"
- *Literacy* Rather than examine mastery of specific school curricula, PISA looks at students' ability to apply knowledge and skills in key subject areas and to analyse, reason and communicate effectively as they examine, interpret and solve problems.
- *Lifelong learning* PISA also asks students about their motivations, beliefs about themselves and learning strategies.

The growing reach...

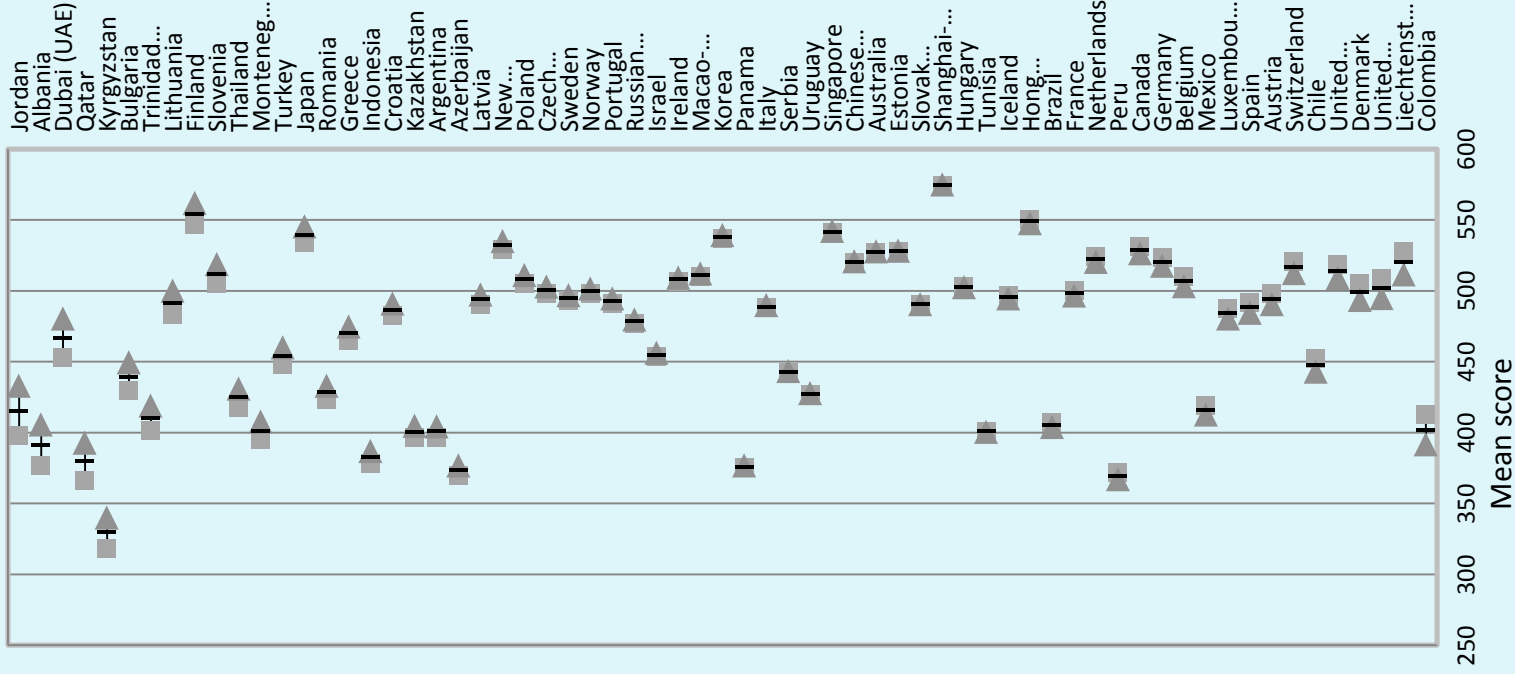
- More powerful analyses:

PISA has created huge amounts of big data about the quality of schooling outcomes. PISA has also helped to change the balance of power in education by making public policy in the field of education more transparent and more efficient. Andreas Schleicher, OECD, July 2013

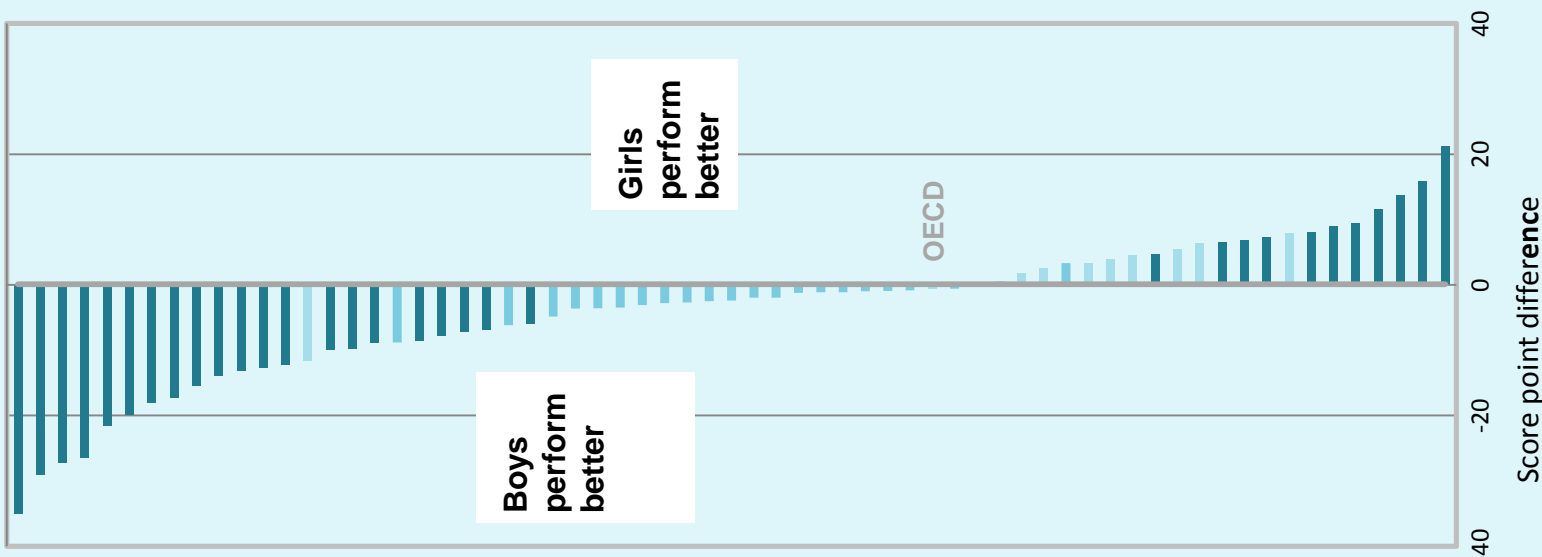
- More countries taking part
- Detailed country analyses
- PISA spin offs, aimed at improving international understanding of educational effectiveness
- Resulting in more countries using PISA to drive their policies (eg 'closing the gap' in the UK, curriculum design in Germany)



Mean score on the science scale



Gender difference (girls - boys)



Five volumes of PISA 2009 products

- *‘What students know and can do – student performance in reading, mathematics and science’*
- *‘Overcoming social background: equity in learning opportunities and outcomes’*
- *‘Learning to learn’*
- *‘What makes a school successful?’*
- *‘Learning trends: changes in student performance since 2000’*

Plus online database of results, assessment framework and sample questions (*‘Take the test’*)

Denial, acceptance and welcome

Only five countries in a 2011 survey reported PISA as having had little or no impact on national policy (reported in OECD Working paper 71, 2012)

- Germany – ‘PISA shock’ in 2000 led to reform of curriculum and action to close performance gaps
- Denmark – heart searching over social equity following 2000 PISA round
- Japan – decline in performance in 2003 led to tightening of national curriculum and assessment system
- UK – relatively poorer 2009 results used to justify controversial school reforms
- Wales – wholesale revision of school improvement strategies after 2009 results
- Finland and Shanghai – outliers or examples to follow?
- And what about Singapore? Are there any lessons to learn? Yes: *“examples of Finland and Shanghai in supporting weak performers or weak schools are instructive as we review our own strategies”* (response to 2011 survey).

Which areas of PISA policy analysis have been influential in national policy-making processes?

a. Assessment and accountability	29
b. Learning environment	13
c. Early childhood education	13
d. Resource invested and allocation	12
e. Student selection and tracking	11
f. Governance (e.g. autonomy, choice, private/public).	11

Typical 'accountability' responses to PISA

- Curriculum reform
- Strengthened national assessment systems, often modelled on PISA
- Introduction of performance targets at national and/or school level
- More rigorous inspection and evaluation regimes

Use of PISA to evaluate reforms

“Along with other studies, PISA is used to provide an indication of the effectiveness of our initiatives to promote critical and inventive thinking; help under-achievers; and maximise the potential of students.”
Response from Singapore to 2011 survey

“PISA is important in monitoring the massive educational reform which started in September 1999 on ISCED 1 and 2 level and in 2001 for ISCED 3 level.” Response from Poland to 2011 survey

Conclusion

- PISA now represents the ‘global standard’
- Used in over 65 countries already, more in the pipeline
- Increasingly used as a source of data for second level policy analysis at national level
- Has opened the door wide for countries to learn from one another

...and now, PISA for schools

The PISA-based test for schools

- *'a student assessment tool geared for use by schools and networks of schools to support research, benchmarking and school improvement efforts'*
- Results calibrated on the Pisa performance scales (7 point scale in Reading, 6 point scale in mathematics and science)
- Different assessments from PISA, but based on the same assessment frameworks
- Designed to yield results at school level, not just national level (so no sampling design)
- Provides information on how different factors within and outside school associate with student performance
- Guidelines governing the proper and improper use of the assessments

Ethical position

‘The PISA-based test for schools is intended to be used for research, benchmarking and school improvement purposes. It is not intended as a high-stakes assessment or for accountability purposes’

But there's still one piece of the
jigsaw missing....

**iPIPS - an International Study of
Children's Development at the Start of
School and during their First School Year**

Developed by the
Centre for Evaluation and Monitoring
University of Durham, UK

Why iPIPS?

- Need a baseline for PISA, TIMSS and PIRLS, to provide value added data
- Need internationally comparable data for assessing effectiveness of early learning policies and practice
- Excellent psychometric properties – both reliability and predictive validity
- Will provide high quality information both for policy makers and for teaching professionals

Policy Questions

- To what extent are later differences in later outcomes (e.g. on PISA) explained by differences when children start school?
- How do children's developing abilities vary across jurisdictions? How does this relate to differences in pre-school policy?
- How do children progress in their first year of school, and how does this vary across jurisdictions?
- What is the link between social and economic factors and children's development across jurisdictions?
- Can the data help to interpret policies on pre-school provision, school starting age, curriculum, pedagogy, teacher training etc?

What is PIPS?

- A diagnostic assessment of children's cognitive and non-cognitive development as they start school
- Repeated at the end of their first year, to assess progress
- Developed in 1994, has been used in 10 countries, 1M children on database
- Originally paper based, now computer adaptive
- Provides almost immediate feedback to schools, for diagnostic and formative use, based on nationally comparative data

What does PIPS assess?

- **Objective assessment**

- Vocabulary acquisition*

- Early reading (concepts about print, letter and word identification, comprehension)*

- Early mathematics (concepts about mathematics, digit identification, shape identification, simple and complex sums)*

- Phonological awareness (repeat words and identifying rhyming words)*

- General cognitive function (short term memory)*

- **Ratings**

- Personal, social and emotional development*

- Behaviour (Inattentiveness, hyperactivity and impulsiveness)*

Assessment with the child

- Computer adaptive test – 20 minutes with a teacher or researcher
- Simple and engaging graphics
- Friendly audio cues
- Stopping rules to prevent child becoming discouraged
- Efficient and accurate measurement against 11 sub-scales
- ‘One year on’ assessment starts from where child reached on previous assessment

Ideas About Reading

PIPS Baseline for England 2010 - 2011 v3.1.97



Previous Test

Back to Start

Replay Audio

Right

Wrong

Can you show me someone who is writing?

Abdul and Lisa looked at the cat.
It was stuck in the tree.



Back an Item

Back to Start

Replay Audio

Right

Wrong

If I want to read this story where should I start?

Reading

The dog has got a red ball.



Previous Test

Back to Start

Replay Audio

Number |

ENTER

Enter no. of words read by child. Read any they can't (7)

C

Previous Test

Back to Start

Replay Audio

Right

Wrong

Do you know what this letter is?

Rhymes

PIPS Baseline for England 2009 - 2010 v3.0.51



Back an Item

Back to Start

Replay Audio

Right

Wrong

Which word sounds the same as mouse?

Ideas About Maths

PIPS Baseline for England 2010 - 2011 v3.1.97



Back an Item

Back to Start

Replay Audio

Right

Wrong

Who is the tallest?



Previous Test

Back to Start

Replay Audio

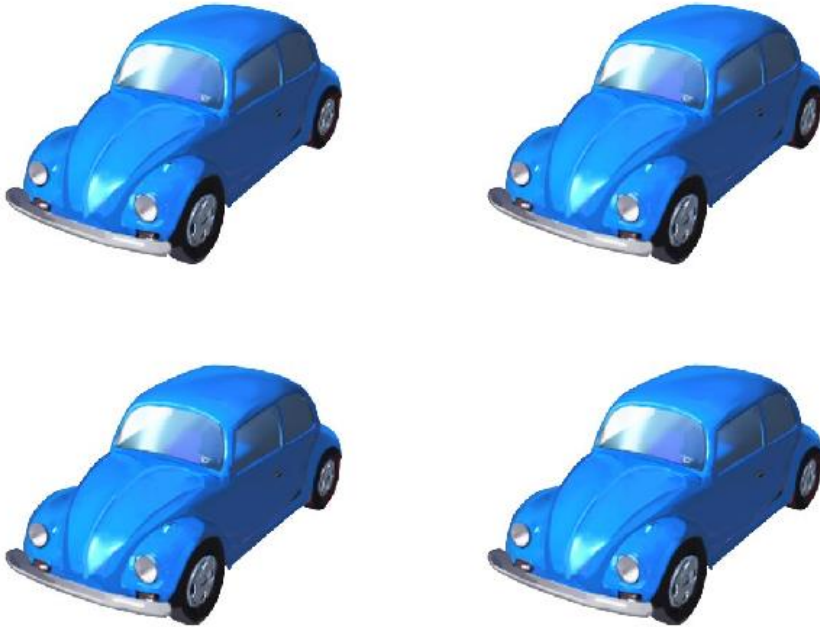
Right

Wrong

Can you show me the biggest cat?

Subtraction

PIPS Baseline for England 2009 - 2010 v3.0.51



Back an Item

Back to Start

Replay Audio

Right

Wrong

How many would be left?

Start generic Microsoft PowerPoint - [...] PIPS Baseline for Engl... 14:44

PIPS Assessment

$$42 - 17 =$$

Back an Item

Back to Start

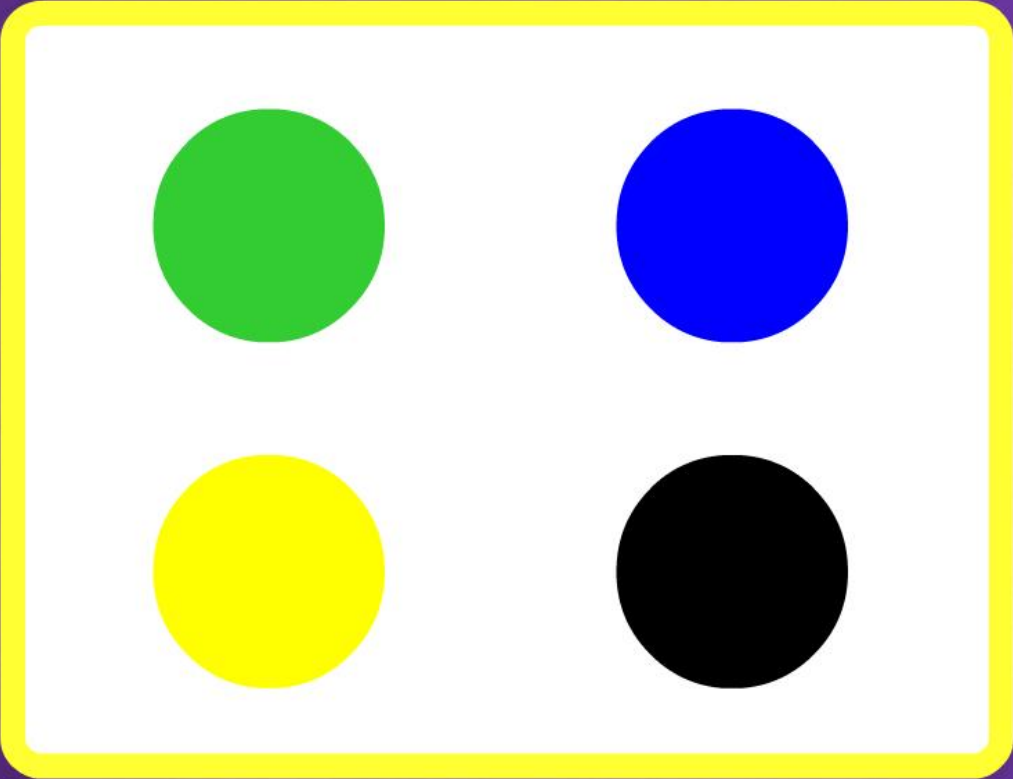
Replay Audio

Right

Wrong

Can you do this?

Executive functioning – short term memory



Back to Start


Repeat

Watch carefully. You are going to see some colours.

The image shows a 2x2 grid of colored circles. The top-left circle is green, the top-right is blue, the bottom-left is yellow, and the bottom-right is black. To the right of the grid are two yellow buttons with black text: 'Back to Start' and 'Repeat'. Below the grid is a yellow rounded rectangle containing the text 'Watch carefully. You are going to see some colours.'

Attitudes

PIPS Baseline for England 2009 - 2010 v3.0.51



Back an Item

Back to Start

Replay Audio

Do you like listening to stories?

Start generic Microsoft PowerPoint - [...] PIPS Baseline for Engl... 14:48

The image shows a software interface for a listening test. At the top, the title bar reads 'PIPS Baseline for England 2009 - 2010 v3.0.51'. The main area has a purple background. A large yellow-bordered box contains three circular icons: a yellow smiley face, a black neutral face, and a black sad face. To the right of this box are three yellow buttons with black text: 'Back an Item', 'Back to Start', and 'Replay Audio'. Below the yellow box is a white rounded rectangle containing the question 'Do you like listening to stories?'. At the bottom, the Windows taskbar is visible, showing the Start button, a 'generic' folder, a 'Microsoft PowerPoint - [...]' window, and the 'PIPS Baseline for Engl...' window. The system tray on the right shows the time as 14:48.

Teacher questionnaire

Assessment

PERSONAL - Concentration

Teacher-directed activities

- 1 Finds it extremely difficult to concentrate. Very rarely settles to one thing and very easily distracted.
- 2 Short concentration span. Finds it difficult to settle down to one thing. Easily distracted.
- 3 Able to settle to a task and concentrate for a sustained period. May be distracted.
- 4 Attends quite well. Able to maintain concentration and is not disturbed by mild distractions.
- 5 Can focus attention, even in the face of competing activities. Has been seen to concentrate for a long period (e.g. 15 minutes).

[Back an Item](#)

[Back to Start](#)

Analysis: What children know and can do

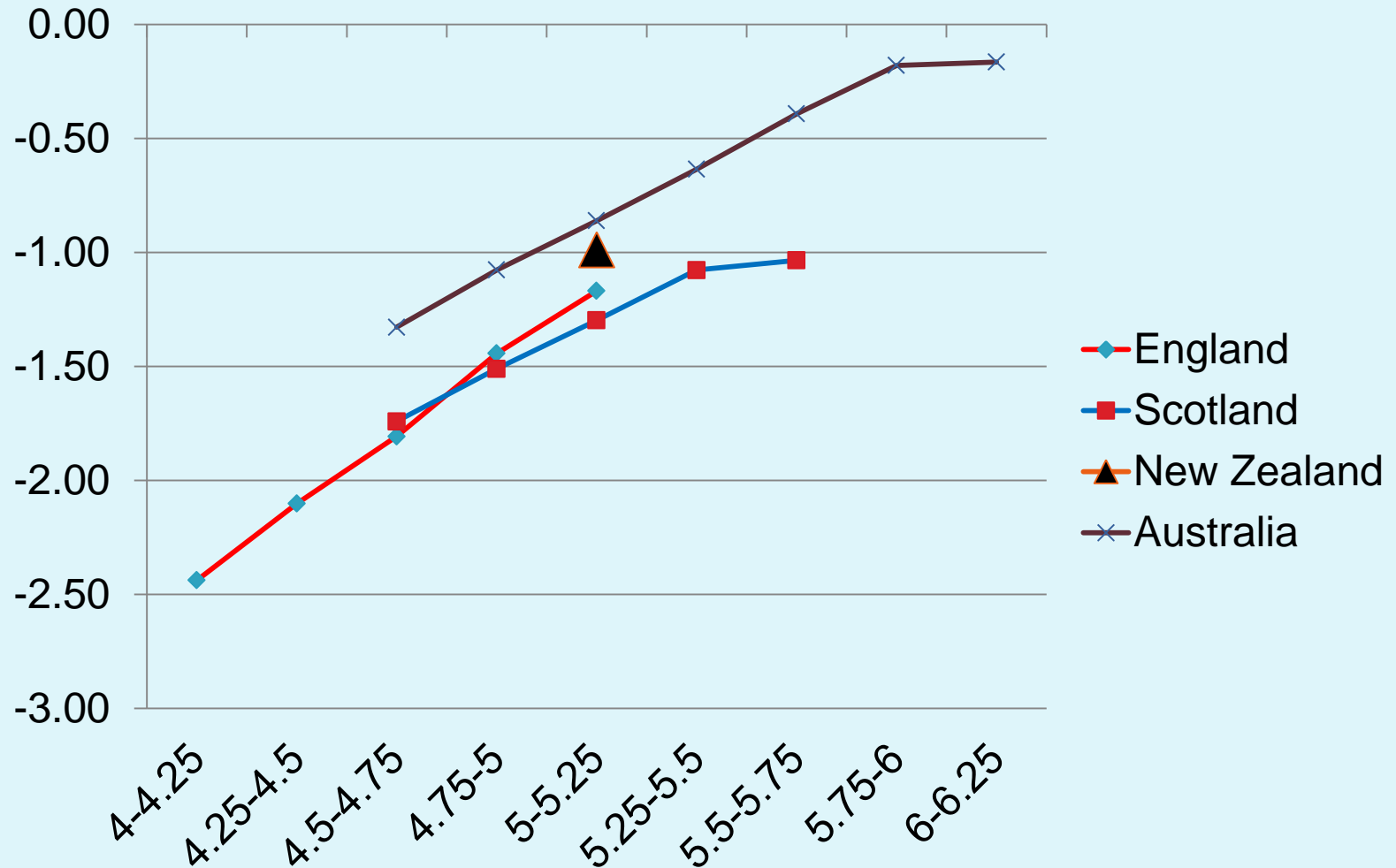
Logits	Distribution of children	Map of items	Difficult
6		+ 42-17=? What is 21 more than 32?	
5	.	+ What is 8 more than 13? What is a quarter of 8? 15+21=? What is half of six?	
4	.	+ 9-6=? 7+3=?	
3	.	+ What is 3 more than 8? Point to some cosmetics.	
2	.# .## .### .#### .#####	+ Read simple sentences, e.g. 'The cat went for a walk'. Identify several two-digit numbers. Read high-frequency words, e.g. dog, tree. Point to a capital letter.	
1	.##### .##### .##### .#####	+ Point to a microscope. Identify all letters. Point to a hexagon.	
0	.##### M+M .##### .##### .##### .#####	+ Identify approximately half of letters. Do informally presented subtraction problems. Repeat 3-syllable words correctly. Identify all single digits.	
-1	.##### .### .##	+ Understand meaning of maths concepts such as 'most' and 'least'. Point to first letter of his/her first name. Detect some rhyming words. Count to 7 and recall counting 7 objects.	
-2	.# .#	+ Identify half of single digits. Count to 4.	
-3	.	+ Point to some cherries. Point to a kite.	
-4	.	+ Understand the meaning of maths concept of 'smallest'. Point to someone writing and someone reading.	
-5	.	+ Point to a fork. Point to some carrots.	

Easy

Using PIPS to compare
children's progress in four
countries

Reading Development on entry

(Illustrative data– not fully representative)



Using PIPS to evaluate the Northern Ireland 'enriched curriculum' on children's acquisition of reading and maths skills

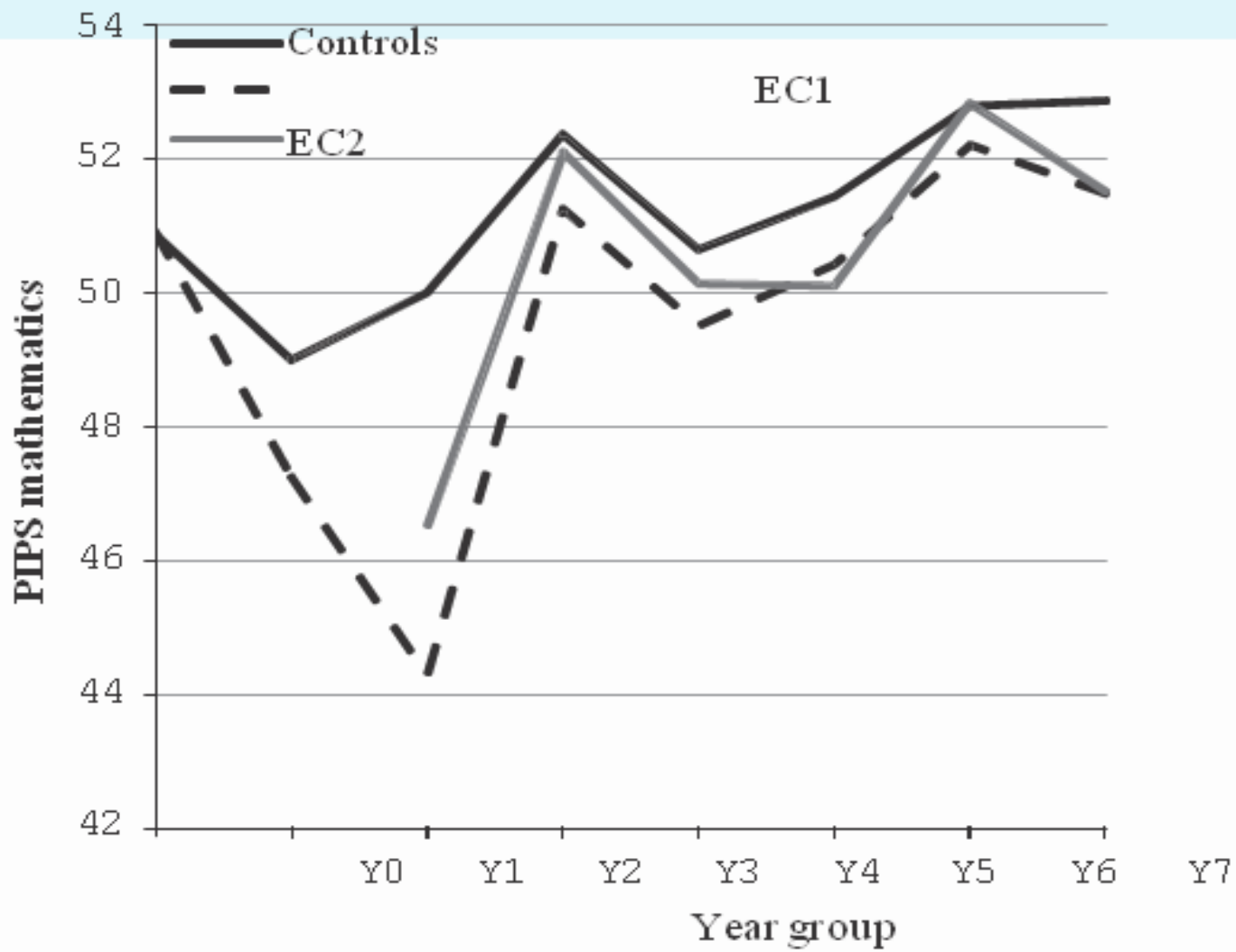


Figure 1. Regression-model-corrected mean scores for mathematics over time

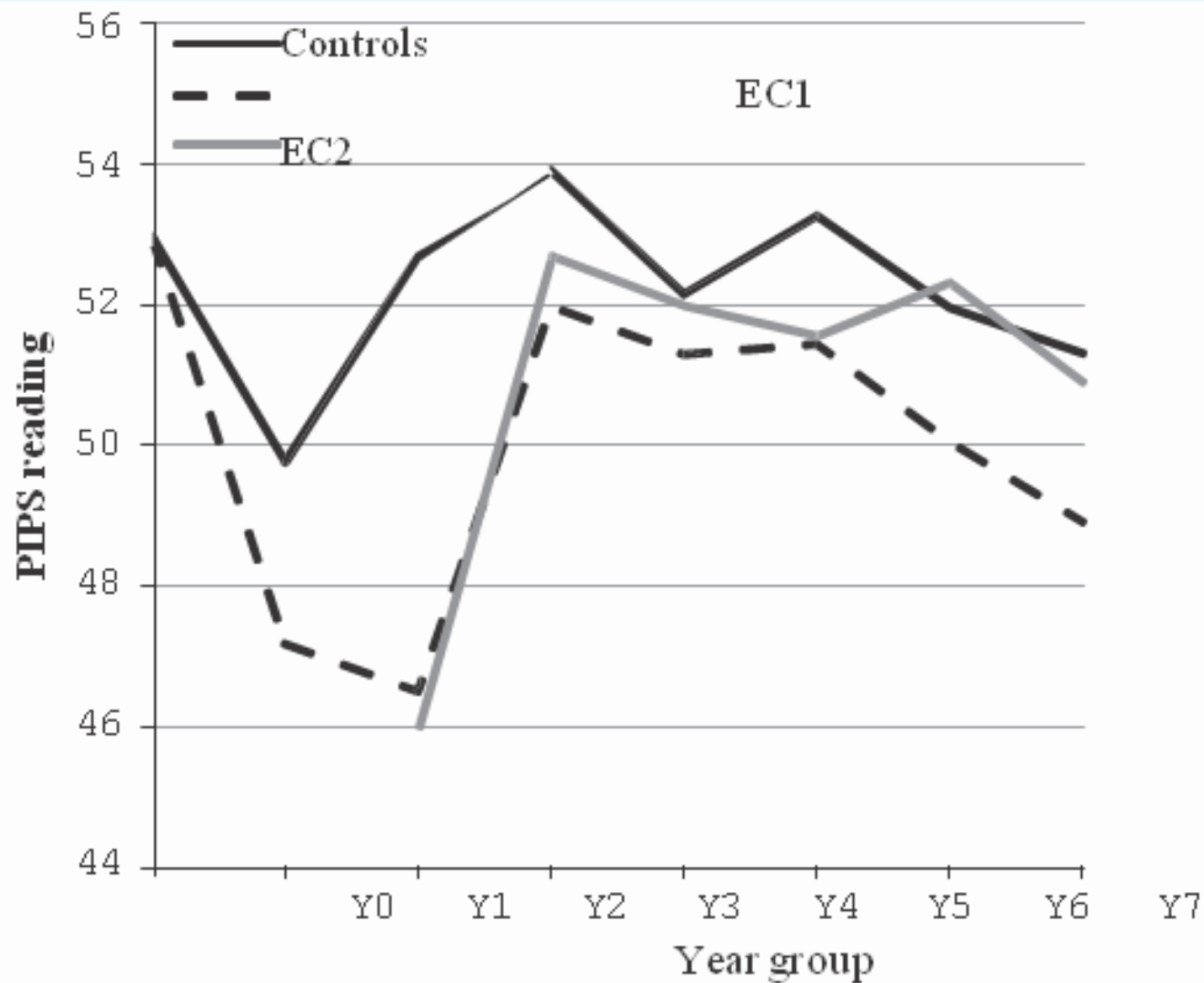


Figure 2. Regression-model-corrected mean scores for reading over time

iPIPS: What is Planned

- Adapt existing PIPS assessment specifically for international comparative use
- Sample based monitoring of c3000 children's developing abilities at start and end of first year in school per country/region
- International and country/regional analyses
- Data for schools to use diagnostically (not accountability or performance management)
- Pilots in 6-8 countries 2013-15
- To be offered more widely thereafter

The iPIPS team - international partner organisations

- Educational Testing Services, US and Worldwide
- Australian Council for Educational Research
- University of Western Australia
- University of Würzburg, Germany
- Centre for Evaluation and Monitoring, Hong Kong
- Centre for Evaluation and Assessment, University of Pretoria, South Africa
- Centre for Evaluation and Monitoring , University of Christchurch, New Zealand
- Higher School of Economics, Moscow
- NIE Singapore and Singapore Principals Academy (hopefully!)

Thank you for your
attention