TABLE OF CONTENTS

COMPARISONS OVER TIME ON THE PISA SCALES 2

PISA trend scales 2
PISA sub-scales 5
Interim scales 5
Background information concerning the construction of the PISA literacy scales 6
Framework Development 6
Testing Time and Item Characteristics 7
Characteristics of each of the links 8
Reading 8
Mathematics 9
Science 9
Use of individual country data for the purpose of trend reporting 10
References 11

COMPARISONS OVER TIME ON THE PISA SCALES

1. PISA 2006 is the third PISA assessment and is therefore the third occasion on which students' scores in reading, mathematics and science have been reported. One of the main aims of PISA is to monitor trends over time. This document sets out the range of cognitive scales that have been prepared over the past three PISA assessments, and describes their special features and appropriate use. It also gives the relevant background information on the PISA test design and characteristics of the links between two cycles in each of the domains. Finally, it reports on important caveats in the use of individual country data for the purpose of trends. The document is based on the OECD document EDU/PISA/GB(2007)42 (Discussion paper for establishing a strategy for the involvement of non-OECD countries in PISA) and on the initial international report of PISA 2006 (OECD, 2007).

2. Table 1 provides a listing of the 19 distinct cognitive scales that were produced as part of PISA assessments in 2000, 2003 and 2006. [1] For the purpose of this overview, the cognitive scales are classified under three different categories: *PISA trend scales*, *PISA sub-scales* and *interim scales*. *PISA trend scales* are the key reporting scales established for each domain, when that domain has been the major domain. *PISA sub-scales* are sub-components of *PISA trend scales* that were provided when a domain was the major domain. *Interim scales* are scales that were temporarily used prior to the establishment of the *PISA trend scales.*

3. In the table, each scale is named, the database upon which it was established is given, the datasets for which it is provided are indicated and comments are made about the scale's appropriate use. The following text provides a more in-depth description of each of these scales.

PISA trend scales

4. The primary PISA reporting scales used for the reporting of trends are the combined scales for reading, mathematics and science. These scales were established in the year in which their respective domain was the major domain, since in that year the framework for the domain was fully developed and the domain was comprehensively assessed. When the combined scale is established, the mean of the scale is set at 500 and the standard deviation is set at 100 (for the pooled, equally weighted OECD countries). For example, 500 on the PISA mathematics combined scale is the mean achievement of the assessed students in OECD countries in 2003.

5. The intention for these trend scales is to stay in place until the specification of the domain is changed or updated.

[1] Note that this section refers to cognitive scales only. PISA has also produced a wide range of other scales that are affective or behavioural scales.

Table 1: Summary of PISA cognitive reporting scales

Name	Established	2000	2003	2006	Comment
PISA trend scales					
PISA Reading	2000	✓	✓	✓	Trends can be reported between any of the three cycles, by country or by subgroups within countries
PISA Mathematics	2003		✓	✓	Trends can be reported between 2003 and 2006, by country or by subgroups within countries
PISA Science	2006			✓	Provides the basis for future trend analysis by country or by subgroups within country
PISA sub-scales					
Reading scale: Retrieving information	2000	✓			
Reading scale: Interpreting texts	2000	✓			
Reading scale: Reflection and evaluation	2000	✓			
Mathematics scale: Quantity	2003		✓		
Mathematics scale: Uncertainty	2003		✓		
Mathematics scale: Space & Shape	2003	✓	✓		Established in 2003 and then applied to 2000 with a rescaling (no conditioning). Trends on this sub-scale can be reported for countries, but are not optimal for subgroups within countries.
Mathematics scale: Change &	2003	✓	✓		Established in 2003 and then applied to 2000 with a rescaling (no conditioning). Trends on this sub-

Relationships					scale can be reported for countries, but are not optimal for subgroups within countries.
Science scale: Explaining Phenomena Scientifically	2006			✓	
Science scale: Identifying Scientific Issues	2006			✓	
Science scale: Using Scientific Evidence	2006			✓	
Science scale: Physical Systems	2006			✓	Limited conditioning implemented permitting unbiased estimation by country and by gender. Results for other subgroups are not optimal.
Science scale: Earth & Space systems	2006			✓	Limited conditioning implemented permitting unbiased estimation by country and by gender. Results for other subgroups are not optimal.
Science scale: Living Systems	2006			✓	Limited conditioning implemented permitting unbiased estimation by country and by gender. Results for other subgroups are not optimal.
Interim scales					
Interim Mathematics	2000	✓			
Interim Science	2000	✓	✓		
Science Common Items 2003-2006	2006		✓	✓	Provides a comparison over time using items that were common to PISA 2003 and 2006; the science framework is not fully covered, therefore, this is not a science trend scale.

PISA sub-scales

6. Across the three PISA assessments a total of 13 scales have been prepared and reported. In PISA 2000, three reading aspect-based scales were prepared; in PISA 2003, four mathematics content-based scales were prepared; and in 2006 a total of six science scales were prepared. For a description of the content of the scales, one can refer to the PISA framework publications (OECD, 2006).

7. Typically, the scales are prepared only in the year in which a domain is a major domain, since it is at this time that there are sufficient items in each sub-area to support the reporting of the scales. The one exception to this rule was mathematics, for which the *space and shape* and *change and relationships* scales were reported for the PISA 2000 data as well as the PISA 2003 data. These scales were established in 2003 when mathematics was the major domain, however, they could be applied to the 2000 data because only those two areas of mathematics (*space and shape* and *change and relationships*) had been assessed in PISA 2000, and sufficient common items were available to support the scaling.

8. For the 2000 data, the mathematics scales were prepared using a methodology that permits analysis over assessment cycles at the national level (or at the level of adjudicated regions), but the scales were not optimal for analysis at the level of student sub-groups. This was due to the fact that conditioning variables were not used in the construction of the scales for the PISA 2000 data (see OECD, 2005).

9. For science in PISA 2006, two alternative sets of scales were prepared. The first was a set of three process-based scales and the second was a set of three content-based scales. It is important to note that these are alternative scalings relying on the same test items. As such, it is inappropriate to jointly analyse scales that are selected from the alternative scalings. For example, it would not be meaningful or defensible to correlate or otherwise compare performance on the *physical systems* scale, with performance on the *using scientific evidence* scale.

10. The process-based science scales are suitable for national analyses and sub-group analyses, whereas analysis of the content-based science scales is suitable only at the national level (or at the level of adjudicated regions), and by gender. The content-based scales are not optimal for use at the level of any other student sub-groups. This is because gender was the only conditioning variable used in the construction of the content-based science scales (see OECD, 2008).

11. The metric of all of the PISA scales is set so that scales within a domain can be compared to each other and with the matching primary PISA combined scale. For science, as mentioned above, a comparison across alternative scalings of the same domain (process-based vs. content-based) is not appropriate

Interim scales

12. There are three special purpose scales.

13. An *Interim Mathematics* scale was established and reported in PISA 2000. This scale was prepared to provide an overall Mathematics score, and it used all of the mathematics items that were included in the PISA 2000 assessment. This scale was discontinued in 2003 when

mathematics became the major domain and the more comprehensive PISA mathematics trend scale was established.

14. An *Interim Science* scale was established and reported in PISA 2000. This scale was prepared to provide an overall Science score, and it used all of the science items that were included in the PISA 2000 assessment. The PISA 2003 science data was linked to this scale so that the PISA 2003 science results were also reported on the interim science scale. For PISA 2006, this scale was not provided because Science became the major domain and the more comprehensive PISA trend science scale was established.

15. To allow comparisons between science outcomes in 2003 and 2006, a *Science Common Items 2003-2006* scale was prepared. This scale is based on the science items that are common to PISA 2003 and PISA 2006, and can be used to examine trends, on those common items, between 2003 and 2006. The PISA 2003 abilities that are based on the common items can be analysed at the national level (or at the level of adjudicated regions) and by gender, but they are not optimal for use at the level of any other student sub-groups. The PISA 2006 abilities, associated with the fully developed PISA combined science scale, can be analysed by national subgroups as well.

Background information concerning the construction of the PISA literacy scales

17. When an assessment area is a major domain, there are two key characteristics that distinguish it from a minor domain. First, the framework for the assessment area is fully developed and elaborated. Second, more assessment time is allocated to the major domain than is allocated to each of the minor domains because the framework is comprehensively assessed.

Framework Development

PISA 2000

18. For PISA 2000, a full and comprehensive framework was developed for reading to guide the assessment of reading as a major domain. Less fully articulated frameworks were developed to support the assessment of mathematics and science as minor domains (OECD, 1999).

PISA 2003

19. For PISA 2003, the mathematics framework was updated and fully developed to support a comprehensive assessment of mathematics. The reading and science frameworks were retained largely as they had been for PISA 2000 (see OECD, 2003).

20. The key changes to the mathematics framework between 2000 and 2003 included:

- The addition of a theoretical underpinning of the mathematics assessment, expanding the rationale for the PISA emphasis on *the use of* mathematical knowledge and skills to solve problems encountered in life.
- The restructuring and expansion of domain content: expansion from two broad content areas (*overarching ideas)* to four; removal of all references to mathematics curricular strands as a separate content categorisation (instead, definitions of the overarching ideas

were expanded to include mention of the *kinds* of school mathematics topics associated with each definition).

- A more elaborate rational for the existing balance between "realistic mathematics" and more traditional context-free items, in line with the "literacy for life" notion underlying OECD/PISA assessments.
- A redeveloped discussion of the relevant mathematical processes: a clearer and much more enhanced link between the process referred to as 'mathematisation', the underlying mathematical competencies, and the competency clusters; and a better operationalisation of the competency classes through a more detailed description of the underlying proficiency demands placed on students.
- Considerable elaboration through the addition of examples, including items from previous test administrations.

21. Clearly, the framework change involving an effective doubling of the mathematical content base of the study was of such significance that trend measures would be very seriously affected. As a result, only sub-scale links to 2000 were possible, and the new framework provided the first comprehensive basis for the calculation of future trend estimates.

PISA 2006

22. For PISA 2006, science became the major domain, so the science framework was updated and fully developed to support a comprehensive assessment of science. The reading framework was retained largely as it had been for PISA 2000, and the mathematics framework as it had been for PISA 2003 (see OECD, 2006). Between 2003 and 2006, key changes to the science framework included:

- A clearer separation between *knowledge about science* as a form of human enquiry and *knowledge of science* as a knowledge of the natural world as articulated in the different scientific disciplines. In particular, PISA 2006 gives greater emphasis to knowledge about science as an aspect of science performance through the addition of elements that underscore students' knowledge about the characteristic features of science and scientific endeavours; and
- The addition of new components about the relationship between science and technology.

23. Both of these changes carry the potential to disrupt links with the previous special purpose science scales: the interim science and trend science scales.

Testing Time and Item Characteristics

24. In each of PISA 2000, 2003 and 2006, a total of 390 unique minutes of testing material were used. The distribution of the testing minutes is provided in Table 2. When a domain is assessed as a major domain, more minutes are devoted to it than for minor domains. For example, 270 minutes were assigned to reading material in PISA 2000 to allow full coverage of the framework. Similarly, PISA 2003 included 210 minutes of mathematics material and PISA 2006 included 210 minutes of science material. When a domain is assessed as a minor domain, the assessment is far

less comprehensive and does not provide an in-depth assessment of the full framework. The full framework is developed when a domain is a major domain.

Table 2: Number of item minutes for each domain for each PISA assessment

	Reading	Mathematics	Science	Total
2000	270	60	60	390
2003	60	210	60	330[2]
2006	60	120	210	390

25. The links in terms of number of items in common for successive pairs of assessments are shown in Table 3.

Table 3: Number of common items between successive PISA assessments

	Reading	Mathematics	Science
As Major Domain	129	84	108
Links 2000-2003	28	20	25
Link 2003-2006	28	48	22

Characteristics of each of the links

Reading

2000 to 2003

26. The PISA reading trend scale was established in 2000 on the basis of a fully developed and articulated framework and a comprehensive assessment of that framework. In PISA 2003, a subset of 28 of the 2000 reading items was selected and used. Equating procedures reported in OECD (2005) were then used to report the PISA 2003 data on the established PISA reading scale.

[2] In 2003 the total testing time was also 390 minutes, but 60 minutes of that testing time was allocated to an assessment of Problem Solving skills.

2003 to 2006

27. The items used in PISA 2003 (units and clusters) were used again in order to link the PISA 2006 data to the PISA reading trend scale,

Mathematics

2000 to 2003

28. The mathematics framework that was prepared for PISA 2000 was preliminary and the assessment was restricted to two of the so-called *big ideas – space and shape*, and *change and relationships*. For the PISA 2003 assessment, when mathematics was a major domain, the framework was fully developed and the assessment was broadened to cover the four *overarching ideas – quantity, uncertainty, space and shape*, and *change and relationships*.

29. As the mathematics framework was fully developed for PISA 2003, the PISA mathematics trend scale was developed at that point as well. As PISA 2000 had covered two of the scales, two scales were developed that permitted comparison of performance between 2000 and 2003. These were the *space and shape* scale and the *change and relationships* scale.

2003 to 2006

30. A selection of 48 mathematics items was selected from PISA 2003 and used again in PISA 2006. The 48 items represented 120 minutes of testing time.

Science

2000 to 2003

31. Science was a minor domain in both PISA 2000 and 2003. As such, the assessment on both of these occasions was less comprehensive than it was in 2006, when a more fully articulated framework was developed and more testing time was allotted. There were 25 items that were common to both PISA 2000 and 2003.

2003 to 2006

32. In PISA 2006, science was the major domain, and as such it was comprehensively assessed on the basis of a newly developed and elaborated framework. There were 108 science items used in PISA 2006, compared with 35 in PISA 2003; of these, just 22 items were common to PISA 2006 and PISA 2003, and 14 were common to PISA 2006 and PISA 2000.

33. So, as the first major assessment of science, the PISA 2006 assessment was used to establish the basis for the PISA science trend scale.

34. For the purposes of comparisons between PISA 2003 and PISA 2006, an additional scale was established based upon those items common to both PISA assessments. The international results of this interim scale are provided in the Annex of the initial report (OECD, 2007; p.369-370).

Use of individual country data for the purpose of trend reporting

35. It is important to note that some countries cannot be included in comparisons between PISA 2000, PISA 2003 and PISA 2006 for methodological reasons. Among OECD countries, the *Slovak Republic* and *Turkey* joined PISA in 2003, and thus cannot be included in any comparisons with PISA 2000.

36. The PISA 2000 sample for *the Netherlands* did not meet the PISA response rate standards, so mean scores for the Netherlands were not reported for PISA 2000. Results for the Netherlands are only available from PISA 2003 onwards.

37. In *Luxembourg*, the assessment conditions were substantially changed between the PISA 2000 and PISA 2003 survey with regard to organisational and linguistic aspects in order to improve compliance with OECD standards and to better reflect the national characteristics of the school system. In PISA 2000, students in Luxembourg were given one assessment booklet with the languages of testing chosen by each student one week prior to the assessment. However, it was found that familiarity with the language of assessment was an important barrier for a significant proportion of students in Luxembourg in PISA 2000. In PISA 2003 and PISA 2006, therefore, students were each given two assessment booklets – one in each of the two languages of instruction - and could choose their preferred language immediately prior to the assessment. This provided for assessment conditions that were more comparable to those in countries with only one language of instruction, and resulted in a fairer assessment of the performance of students in mathematics, science, reading and problem solving. As a result of this change in procedures, the assessment conditions, and hence the assessment results for Luxembourg cannot be compared between PISA 2000 and the following PISA assessments. Assessment conditions between PISA 2003 and PISA 2006 have not been changed, and therefore results can be compared.

38. The PISA 2000 and PISA 2003 samples for the *United Kingdom* did not meet the PISA response rate standards. In PISA 2000, the initial response rate for the United Kingdom fell 3.7% short of the minimum requirement. At that time, the United Kingdom provided evidence to the PISA Consortium that permitted an assessment of the expected performance of the non-participating schools on the basis of which the PISA Consortium concluded that the response-bias was likely negligible. Therefore, the results were included in the international report on PISA 2000. In PISA 2003, the United Kingdom's response rate was such that the required sampling standards were not met and further investigation by the PISA Consortium did not confirm that the resulting response bias was negligible. Therefore, the data was not deemed internationally comparable. For PISA 2006, the more stringent standards were again applied, and therefore PISA 2000 and PISA 2003 data for the United Kingdom cannot be included in the comparisons over time. Therefore, no trend data can be reported for the United Kingdom for PISA 2000, PISA 2003 and PISA 2006.

40. For the *United States*, no reading results are available for PISA 2006. Due to an error in printing the test booklets, some of the reading items had incorrect instructions and the mean performance in reading cannot be accurately estimated. The impact of the error on estimates of student performance is likely to exceed one standard error of sampling. For details, see Annex A3 of the PISA 2006 initial international report (OECD, 2007). This was not the case for science and

mathematics items, therefore comparisons of mathematics results from PISA 2003 to PISA 2006 are available for the United States.

41. For *Austria*, there have been modifications to the weighting of their PISA 2000 data. As noted in the Technical Report for PISA 2000 (OECD, 2002), the Austrian sample for the PISA 2000 assessment did not adequately cover students enrolled in combined school and work-based vocational programmes, as required by the technical standards for PISA. The published PISA 2000 estimates for Austria were therefore biased (OECD, 2001). This non-conformity was corrected in the PISA 2003 assessment. To allow reliable comparisons, adjustments and modified student weights were developed, which make the PISA 2000 estimates comparable to those obtained in PISA 2003. OECD Education Working Paper No. 5 "PISA 2000: Sample Weight Problems in Austria" is available at http://www.oecd.org/dataoecd/3/59/36892238.pdf and presents further details on this issue.

References

Adams, R.J. and Wu, M.L. (2002). *PISA 2000 Technical Report.* Paris: OECD Publications.

OECD (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Paris: OECD Publications.

OECD (2001). *Knowledge and Skills for Life – First Results from PISA 2000*. Paris: OECD Publications.

OECD (2003). *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*. Paris: OECD Publications.

OECD (2004). *Learning for Tomorrow's World – First Results from PISA 2003.* Paris: OECD Publications.

OECD (2005). *PISA 2003 Technical Report.* Paris: OECD Publications.

OECD (2006). *Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006.* Paris: OECD Publications.

OECD (2007). *PISA 2006: Science Competencies for Tomorrow's World.* Paris: OECD Publications.

OECD (2008). *PISA 2006 Technical Report*. Paris: OECD Publications.