5

The Rasch Model

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

INTRODUCTION

International surveys in education such as PISA are designed to estimate the performance in specific subject domains of various subgroups of students, at specific ages or grade levels.

For the surveys to be considered valid, many items need to be developed and included in the final tests. The OECD publications related to the assessment frameworks indicate the breadth and depth of the PISA domains, showing that many items are needed to assess a domain as broadly defined as, for example, mathematical literacy.[1]

At the same time, it is unreasonable and perhaps undesirable to assess each sampled student with the whole item battery because:

- After extended testing time, students' results start to be affected by fatigue and this can bias the outcomes of the surveys.
- School principals would refuse to free their students for the very long testing period that would be required. This would reduce the school participation rate, which in turn might substantially bias the outcomes of the results.

To overcome the conflicting demands of limited student-level testing time and broad coverage of the assessment domain, students are assigned a subset of the item pool. The result of this is that only certain subsamples of students respond to each item.

If the purpose of the survey is to estimate performance by reporting the percentage of correct answers for each item, it would not be necessary to report the performance of individual students. However, typically there is a need to summarise detailed item-level information for communicating the outcomes of the survey to the research community, to the public and also to policy makers. In addition, educational surveys aim to explain the difference in results between countries, between schools and between students. For instance, a researcher might be interested in the difference in performance between males and females.

HOW CAN THE INFORMATION BE SUMMARISED?

At the country level, the most straightforward procedure for summarising the item-level information would be to compute the average percentage of correct answers. This has been largely used in previous national and international surveys and is still used in some current international surveys, even when more complex models are implemented. These surveys may report the overall percentage of correct answers in reading, in mathematics and in science, as well as by content areas (for example, biology, physics, chemistry, earth sciences and so on). For instance, in mathematics, the overall percentage of correct answers for one country might be 54%, and for another, 65%.

The great advantage of this type of reporting is that it can be understood by everyone. Everybody can imagine a mathematics test and can envision what is represented by 54% and 65% of correct answers. These two numbers also give a sense of the difference between two countries.

Nevertheless, there are some weaknesses in this approach, because the percentage of correct answers depends on the difficulty of the test. The actual size of the difference in results between two countries depends on the difficulty of the test, which may lead to misinterpretation.

International surveys do not aim to just report an overall level of performance. Over the past few decades, policy makers have also largely been interested in equity indicators. They may also be interested in the amount of dispersion of results in their country. In some countries the results may be clustered around the mean and in other countries there may be large numbers of students scoring very high results and very low results.

It would be impossible to compute dispersion indices with only the difficulty indices, based on percentage of correct answers of all items. To do so, the information collected through the test need also be summarised at the student level.

To compare the results of two students assessed by two different tests, the tests must have exactly the same average difficulty. For PISA, as all items included in the main study are usually field trialled, test developers have some idea of the item difficulties and can therefore allocate the items to the different tests in such a way that the items in each test have more or less the same average difficulty. However, the two tests will never have exactly the same difficulty.

The distribution of the item difficulties will affect the distribution of the students' performance expressed as a raw score. For instance, a test with only items of medium difficulty will generate a different student score distribution than a test that consists of a large range of item difficulties.

This is complicated to a further degree in PISA as it assesses three or even four domains per cycle. This multiple assessment reduces the number of items available for each domain per test and it is easier to guarantee the comparability of two tests of 60 items than it is with, for example, 15 items.

If the different tests are randomly assigned to students, then the equality of the subpopulations in terms of mean score and variance of the student's performance can be assumed. In other words,

- The mean of the raw score should be identical for the different tests.
- The variance of the student raw scores should be identical for the different tests.

If this is not the case, then it would mean that the different tests do not have exactly the same psychometric properties. To overcome this problem of comparability of student performance between tests, the student's raw scores can be standardised per test. As the equality of the subpopulations can be assumed, differences in the results are due to differences in the test characteristics. The standardisation would then neutralise the effect of test differences on student's performance.

However, usually, only a sample of students from the different subpopulations is tested. As explained in Chapters 3 and 4, this sampling process generates an uncertainty around any population estimates. Therefore, even if different tests present exactly the same psychometric properties and are randomly assigned, the mean and standard deviation of the students' performance between the different tests can differ slightly. As the effect of the test characteristics and the sampling variability cannot be disentangled, the assumption cannot be made that the student raw scores obtained with different tests are fully comparable.

Other psychometric arguments can also be invoked against the use of raw scores based on the percentage of correct answers to assess student performance. Raw scores are on a ratio scale insofar as the interpretation of the results is limited to the number of correct answers. A student who gets a 0 on this scale did not provide any correct answers, but could not be considered as having no competencies, while a student who gets 10 has twice the number of correct answers as a student who gets 5, but does not necessarily have twice the competencies. Similarly, a student with a perfect score could not be considered as having all competencies (Wright and Stone, 1979).

THE RASCH MODEL FOR DICHOTOMOUS ITEMS

Introduction to the Rasch Model

PISA applies the Rasch Model for scaling, in order to overcome challenges identified in the previous section. The following section provides a general introduction to the Rasch Model.

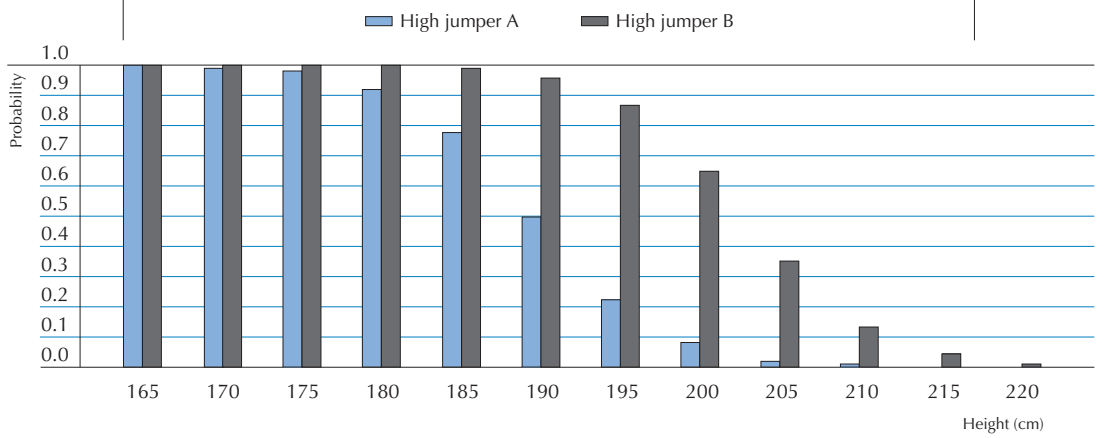
PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

Let us suppose that someone wants to estimate the competence of a high jumper. It might be measured or expressed as his or her:

- individual record,
- individual record during an official and international event,
- mean performance during a particular period of time,
- most frequent performance during a particular period of time.

Figure 5.1 presents the probability of success for two high jumpers per height for the competitions in the previous year.

Figure 5.1
Probability of success for two high jumpers by height (dichotomous)



The two high jumpers always succeeded at 165 centimetres. Then the probability of success progressively decreases to reach 0 for both jumpers at 225 centimetres. While it starts to decrease at 170 centimetres for High jumper A, it starts to decrease at 185 for High jumper B.

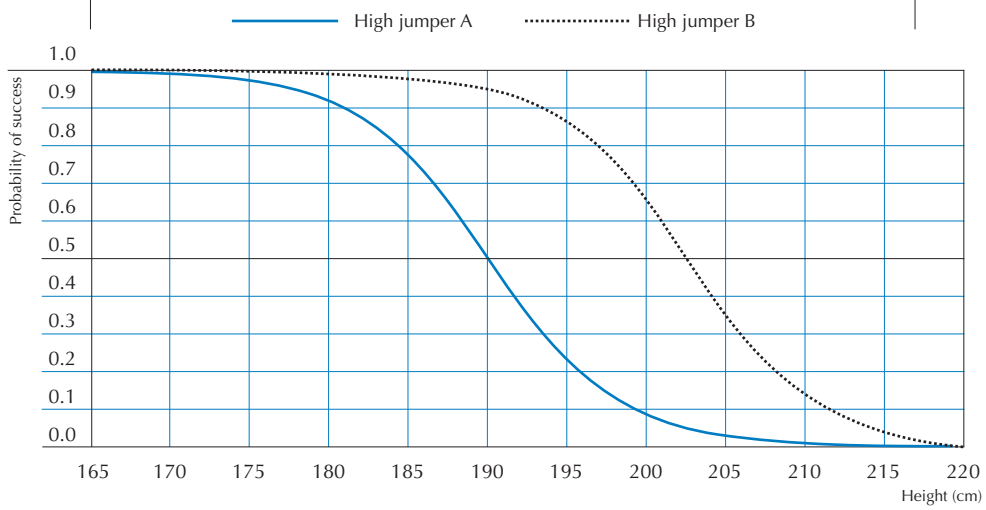
These data can be depicted by a logistic regression model. This statistical analysis consists of explaining a dichotomous variable by a continuous variable. In this example, the continuous variable will explain the success or failure of a particular jumper by the height of the jump. The outcome of this analysis will allow the estimation of the probability of success, given any height. Figure 5.2 presents the probability of success for two high jumpers.

These two functions model the probability of success for the two high jumpers. The blue function represents the probability of success for High jumper A and the black function, the probability of success for High jumper B.

By convention,[2] the performance level would be defined as the height where the probability of success is equal to 0.50. This makes sense as below that level, the probability of success is lower than the probability of failure and beyond that level, it is the inverse.

In this particular example, the performance of the two high jumpers is respectively 190 and 202.5. Note that from Figure 5.1, the performance of High jumper A is directly observable whereas for High jumper B, it needs to be estimated from the model. A key property of this kind of approach is that the level (*i.e.* the height) of the crossbar and the performance of the high jumpers are expressed on the same metric or scale.

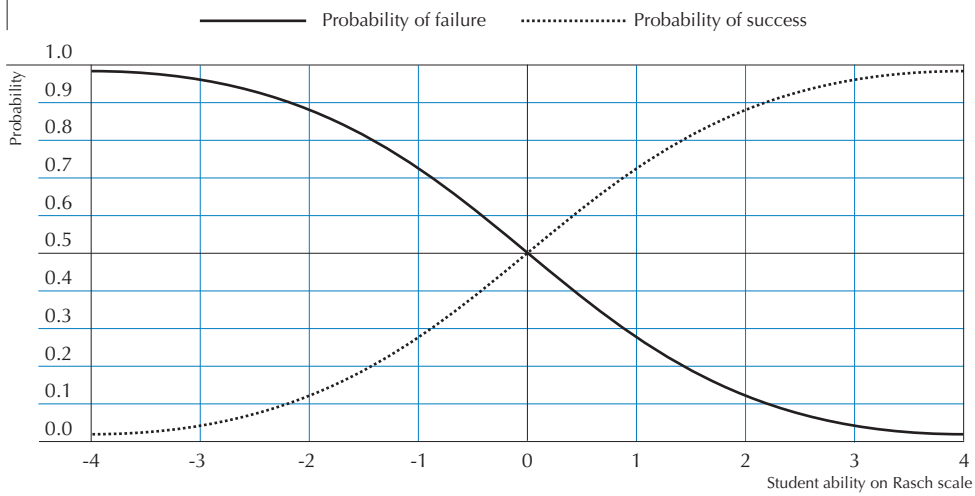
Figure 5.2
Probability of success for two high jumpers by height (continuous)



Scaling cognitive data according to the Rasch Model follows the same principle. The difficulty of the items is analogous to the difficulty of the jump based on the height of the crossbar. Further, just as a particular jump has two possible outcomes, *i.e.* success or failure, a student's answer to a particular question is either correct or incorrect. Finally, just as each jumper's performance was defined at the point where the probability of success was 0.5, the student's performance/ability is measured where the probability of success on an item equals 0.5.

One of the important features of the Rasch Model is that it will create a continuum on which both student performance and item difficulty will be located and a probabilistic function links these two components. Low ability students and easy items will be located on the left side of the continuum or scale, while high ability students and difficult items will be located on the right side of the continuum. Figure 5.3 represents the probability of success and the probability of failure for an item of difficulty zero.

Figure 5.3
Probability of success to an item of difficulty zero as a function of student ability



PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

As shown in Figure 5.3, a student with an ability of zero has a probability of 0.5 of success on an item of difficulty zero and a probability of 0.5 of failure. A student with an ability of –2 has a probability of a bit more than 0.10 of success and a probability of a bit less than 0.90 of failure on the same item of difficulty zero. But this student will have a probability of 0.5 of succeeding on an item of difficulty –2.

From a mathematical point of view, the probability that a student *i*, with an ability denoted β_i, provides a correct answer to item *j* of difficulty δ_j is equal to:

$$P(X_{ij}=1 \mid \beta_i, \delta_j) = \frac{\exp(\beta_i - \delta_j)}{1+\exp(\beta_i - \delta_j)}$$

Similarly, the probability of failure is equal to:

$$P(X_{ij}=0 \mid \beta_i, \delta_j) = \frac{1}{1+\exp(\beta_i - \delta_j)}$$

It can be easily shown that:

$$P(X_{ij}=1 \mid \beta_i, \delta_j) + P(X_{ij}=0 \mid \beta_i, \delta_j) = 1$$

In other words, the probability of success and the probability of failure always sum to one. Tables 5.1 to 5.5 present the probability of success for different student abilities and different item difficulties.

Table 5.1
Probability of success when student ability equals item difficulty

Student ability	Item difficulty	Probability of success
-2	-2	0.50
-1	-1	0.50
0	0	0.50
1	1	0.50
2	2	0.50

Table 5.2
Probability of success when student ability is less than the item difficulty by 1 unit

Student ability	Item difficulty	Probability of success
-2	-1	0.27
-1	0	0.27
0	1	0.27
1	2	0.27
2	3	0.27

Table 5.3
Probability of success when student ability is greater than the item difficulty by 1 unit

Student ability	Item difficulty	Probability of success
-2	-3	0.73
-1	-2	0.73
0	-1	0.73
1	0	0.73
2	3	0.73

Table 5.4
Probability of success when student ability is less than the item difficulty by 2 units

Student ability	Item difficulty	Probability of success
-2	0	0.12
-1	1	0.12
0	2	0.12
1	3	0.12
2	4	0.12

Table 5.5
Probability of success when student ability is greater than the item difficulty by 2 units

Student ability	Item difficulty	Probability of success
-2	-4	0.88
-1	-3	0.88
0	-2	0.88
1	-1	0.88
2	0	0.88

It can be observed that:

- When the student ability is equal to the item difficulty, the probability of success will always be equal to 0.50, regardless of the student ability and item difficulty locations on the continuum (Table 5.1).
- If the item difficulty exceeds the student ability by one Rasch unit, denoted as a logit, then the probability of success will always be equal to 0.27, regardless of the location of the student ability on the continuum (Table 5.2).
- If the student ability exceeds the item difficulty by one logit, the probability of success will always be equal to 0.73, regardless of the location of the student ability on the continuum (Table 5.3).
- If two units separate the student ability and the item difficulty, the probabilities of success will be 0.12 when the student ability is lower than the item difficulty and 0.88 when the student ability is higher than the item difficulty (Tables 5.4 and 5.5).

From these observations, it is evident that the only factor that influences the probability of success is the distance on the Rasch continuum between the student ability and the item difficulty.

These examples also illustrate the symmetry of the scale. If student ability is lower than item difficulty by one logit, then the probability of success will be 0.27, which is 0.23 lower than the probability of success when ability and difficulty are equal. If student ability is higher than item difficulty by one logit, the probability of success will be 0.73, which is 0.23 higher than the probability of success when ability and difficulty are equal. Similarly, a difference of two logits generates a change of 0.38 from the probability of success when ability and difficulty are equal.

Item calibration

Of course, in real settings a student's answer will either be correct or incorrect, so what then is the meaning of a probability of 0.5 of success in terms of correct or incorrect answers? In simple terms the following interpretations can be made:

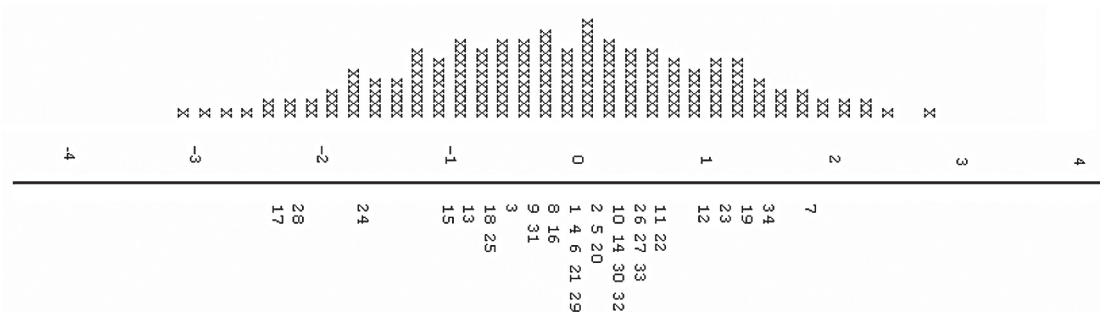
- If 100 students each having an ability of 0 have to answer a item of difficulty 0, then the model will predict 50 students with correct answers and 50 students with incorrect answers.
- If a student with an ability of 0 has to answer 100 items, all of difficulty 0, then the model will predict 50 correct answers and 50 incorrect answers.

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

As described, the Rasch Model, through a probabilistic function, builds a relative continuum on which the item's difficulty and the student's ability are located. With the example of high jumpers, the continuum already exists, *i.e.* this is the physical continuum of the meter height. With cognitive data, the continuum has to be built. By analogy, this consists of building a continuum on which the unknown height of the crossbars, *i.e.* the difficulty of the items, will be located. The following three major principles underlie the construction of the Rasch continuum.

- The relative difficulty of an item results from the comparison of that item with all other items. Let us suppose that a test consists of only two items. Intuitively, the response pattern (0, 0) and (1, 1) (1 denotes a success and 0 denotes a failure), where the ordered pairs refer to the responses to items 1 and 2, respectively, is uninformative for comparing the two items. The responses in these patterns are identical. On the other hand, responses (1, 0) and (0, 1) are different and are informative on just that comparison. If 50 students have the (0, 1) response pattern and only 10 students have the (1, 0) response pattern, then the second item is substantially easier than the first item. Indeed, 50 students succeeded on the second item while failing the first one and only 10 students succeeded on the first item while failing the second. This means that if one person succeeds on one of these two items, the probability of succeeding on the second item is five times higher than the probability of succeeding on first item. It is, therefore, easier to succeed on the second than it is to succeed on the first. Note that the relative difficulty of the two items is independent of the student abilities.
- As difficulties are determined through comparison of items, this creates a relative scale, and therefore there are an infinite number of scale points. Broadly speaking, the process of overcoming this issue is comparable to the need to create anchor points on the temperature scale. For example, Celsius fixed two reference points: the temperature at which the water freezes and the temperature at which water boils. He labelled the first reference point as 0 and the second reference point at 100 and consequently defined the measurement unit as one-hundredth of the distance between the two reference points. In the case of the Rasch Model, the measurement unit is defined by the probabilistic function involving the item difficulty and student ability parameters. Therefore, only one reference point has to be defined. The most common reference point consists of centring the item difficulties on zero. However, other arbitrary reference points can be used, like centring the student's abilities on zero.
- This continuum allows the computation of the relative difficulty of items partly submitted to different subpopulations. Let us suppose that the first item was administered to all students and the second item was only administered to the low ability students. The comparison of items will only be performed on the subpopulation who was administered both items, *i.e.* the low ability student population. The relative difficulty of the two items will be based on this common subset of students.

Figure 5.4
Student score and item difficulty distributions on a Rasch continuum



PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

Once the item difficulties have been placed on the Rasch continuum, the student scores can be computed. The line in Figure 5.4 represents a Rasch continuum. The item difficulties are located above that line and the item numbers are located below the line. For instance, item 7 represents a difficult item and item 17, an easy item. This test includes a few easy items, a large number of medium difficulty items and a few difficult items. The x symbols above the line represent the distribution of the student scores.

Computation of a student's score

Once the item difficulties have been located on the Rasch scale, student scores can be computed. As mentioned previously, the probability that a student *i*, with an ability denoted β_i, provides a correct answer to item *j* of difficulty δ_j is equal to:

$$P(X_{ij}=1 \mid \beta_i, \delta_j) = \frac{\exp(\beta_i - \delta_j)}{1+\exp(\beta_i - \delta_j)}$$

Similarly, the probability of failure is equal to:

$$P(X_{ij}=0 \mid \beta_i, \delta_j) = \frac{1}{1+\exp(\beta_i - \delta_j)}$$

The Rasch Model assumes the independence of the items, *i.e.* the probability of a correct answer does not depend on the responses given to the other items. Consequently, the probability of succeeding on two items is equal to the product of the two individual probabilities of success.

Let us consider a test of four items with the following items difficulties: –1, –0.5, 0.5 and 1. There are 16 possible responses patterns. These 16 patterns are presented in Table 5.6.

Table 5.6
Possible response pattern for a test of four items

Raw score	Response patterns
0	(0, 0, 0, 0)
1	(1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1)
2	(1, 1, 0, 0), (1, 0, 1, 0), (1, 0, 0, 1), (0, 1, 1, 0), (0, 1, 0, 1), (0, 0, 1, 1)
3	(1, 1, 1, 0), (1, 1, 0, 1), (1, 0, 1, 1), (0, 1, 1, 1)
4	(1, 1, 1, 1)

For any student ability denoted β_i, it is possible to compute the probability of any response pattern. Let us compute the probability of the response pattern (1, 1, 0, 0) for three students with an ability of -1, 0, and 1.

Table 5.7
Probability for the response pattern (1, 1, 0, 0) for three student abilities

			$\beta_i = -1$	$\beta_i = 0$	$\beta_i = 1$
Item 1	$\delta_1 = -1$	Response = 1	0.50	0.73	0.88
Item 2	$\delta_2 = -0.5$	Response = 1	0.38	0.62	0.82
Item 3	$\delta_3 = 0.5$	Response = 0	0.82	0.62	0.38
Item 4	$\delta_4 = 1$	Response = 0	0.88	0.73	0.50
Probability of obtaining response pattern			**0.14**	**0.21**	**0.14**

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

The probability of success for the first student on the first item is equal to:

$$P(X_{ij}=1\,|\,\beta_i,\delta_j)=P(X_{1,1}=1\,|\,-1,-1)\frac{\exp(-1-(-1))}{1+\exp(-1-(-1))}=0.5$$

The probability of success for the first student on the second item is equal to:

$$P(X_{ij}=1\,|\,\beta_i,\delta_j)=P(X_{1,2}=1\,|\,-1,-0.5)\frac{\exp(-1-(0.5))}{1+\exp(-1-(-0.5))}=0.38$$

The probability of failure for the first student on the third item is equal to:

$$P(X_{ij}=0\,|\,\beta_i,\delta_j)=P(X_{1,3}=0\,|\,-1,0.5)\frac{1}{1+\exp(-1-0.5)}=0.82$$

The probability of failure for the first student on the fourth item is equal to:

$$P(X_{ij}=0\,|\,\beta_i,\delta_j)=P(X_{1,4}=0\,|\,-1,1)\frac{1}{1+\exp(-1-1)}=0.88$$

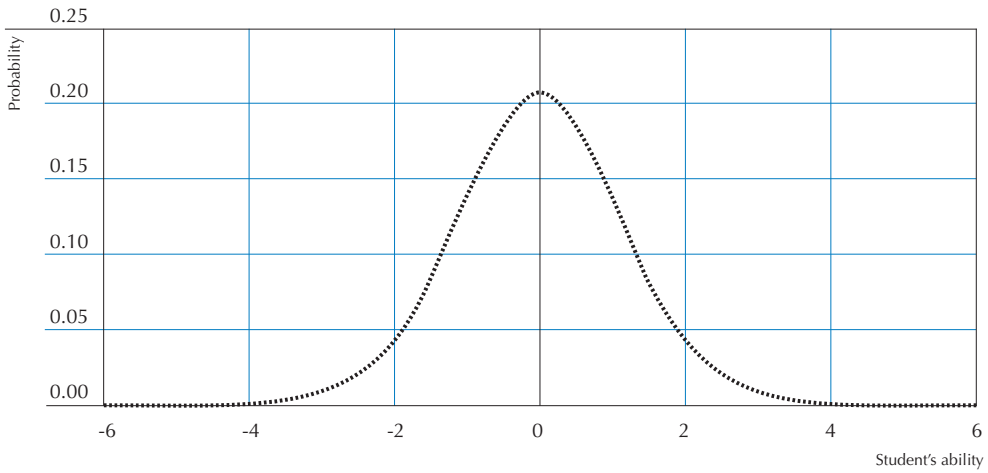
As these four items are considered as independent, the probability of the response pattern (1, 1, 0, 0) for a student with an ability $\beta_i = -1$ is equal to:

0.50*0.38*0.82*0.88 = 0.14

Given the item difficulties, a student with an ability $\beta_i = -1$ has 14 chances out of 100 to provide a correct answer to items 1 and 2 and to provide an incorrect answer to items 3 and 4. Similarly, a student with an ability of $\beta_i = 0$ has a probability of 0.21 to provide the same response pattern and a student with an ability of $\beta_i = 1$ has a probability of 0.14.

This process can be applied for a large range of student abilities and for all possible response patterns. Figure 5.5 presents the probability of observing the response pattern (1, 1, 0, 0) for all students' abilities between –6 and +6. As shown, the most likely value corresponds to a student ability of 0. Therefore, the Rasch Model will estimate the ability of any students with a response pattern (1, 1, 0, 0) to 0.

Figure 5.5
Response pattern probabilities for the response pattern (1, 1, 0, 0)



PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

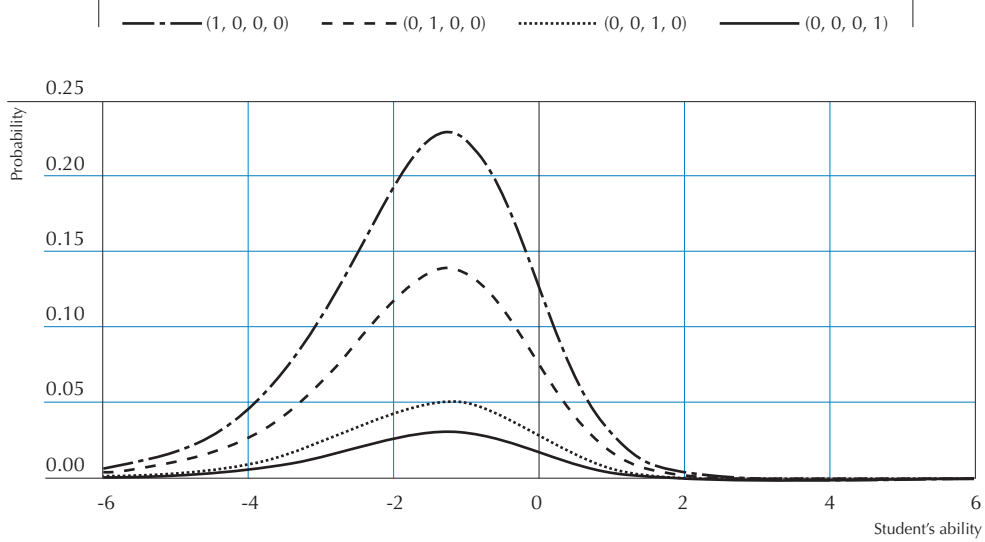
Figure 5.6 presents the distribution of the probabilities for all response patterns with only one correct item. As shown in Table 5.6, there are four responses patterns with only one correct item, *i.e.* (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1).

Figure 5.6 clearly shows that:

- The most likely response pattern for any students who succeed on only one item is (1, 0, 0, 0) and the most unlikely response pattern is (0, 0, 0, 1). When a student only provides one correct answer, it is expected that the correct answer was provided for the easiest item, *i.e.* item 1. It is also unexpected that this correct answer was provided for the most difficult item, *i.e.* item 4.
- Whatever the response pattern, the most likely value always corresponds to the same value for student ability. For instance, the most likely student ability for the response pattern (1, 0, 0, 0) is around –1.25. This is also the most likely student ability for the other response patterns.

The Rasch Model will therefore return the value –1.25 for any students who get only one correct answer, whichever item was answered correctly.

Figure 5.6
Response pattern probabilities for a raw score of 1

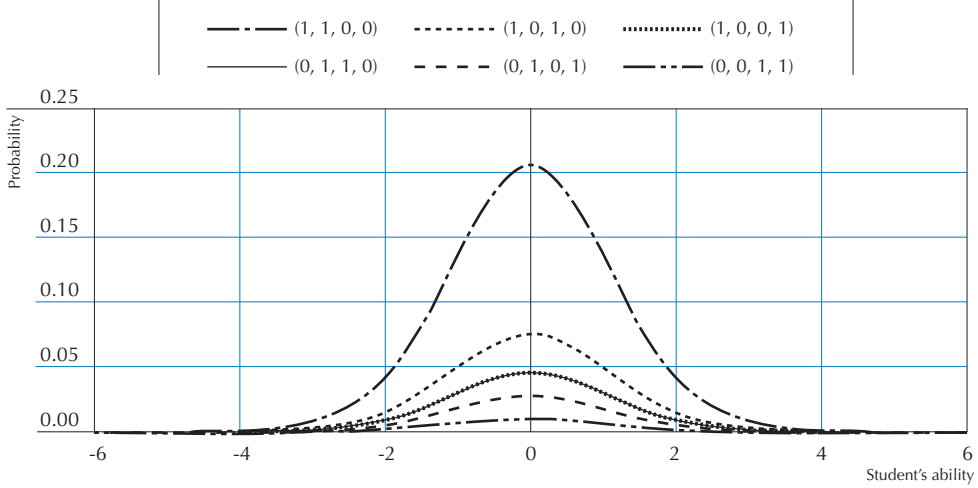


Similarly, as shown by Figure 5.7 and by Figure 5.8:

- The most likely response pattern with two correct items is (1, 1, 0, 0).
- The most likely student ability is always the same for any response pattern that includes two correct answers (student ability is 0 in this case).
- The most likely response pattern with three correct items is (1, 1, 1, 0).
- The most likely student ability is always the same for any response pattern that includes three correct answers (student ability is +1.25 in this case).

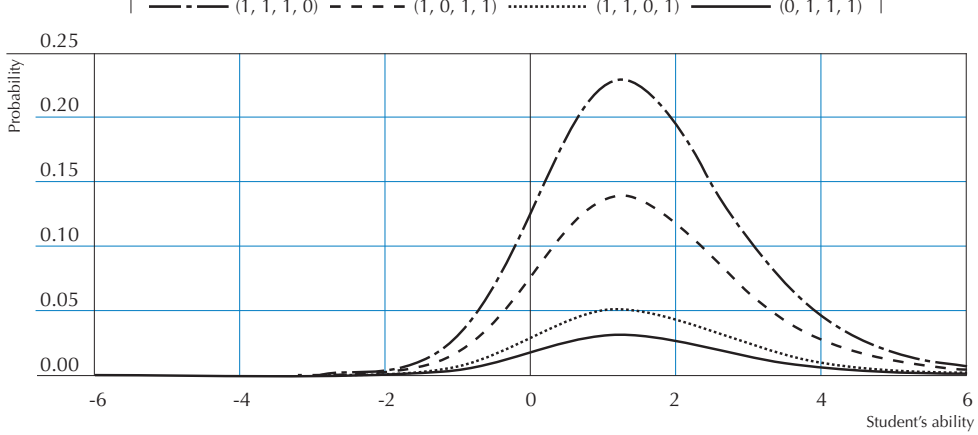
PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

Figure 5.7
Response pattern probabilities for a raw score of 2[a]



a. In this example, since the likelihood function for the response pattern (1, 0, 0, 1) is perfectly similar to that for the response pattern (0, 1, 1, 0), these two lines overlap in the figure.

Figure 5.8
Response pattern probabilities for a raw score of 3



This type of Rasch ability estimate is usually denoted the maximum likelihood estimate (or MLE). As shown by these figures, per raw score, *i.e.* zero correct answers, one correct answer, two correct answers, and so on, the Rasch Model will return only one maximum likelihood estimate.

Warm has shown that this maximum likelihood estimate is biased and proposed to weight the contribution of each item by the information this item can provide (Warm, 1989). Warm estimates and MLEs are similar types of student individual ability estimates.

As the Warm estimate corrects the small bias in the MLE, it is usually preferred as the estimate of an individual's ability. Therefore, in PISA, weighted likelihood estimates (WLEs) are calculated by applying weights to MLE in order to account for the bias inherent in MLE as Warm proposed.

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

Computation of a student's score for incomplete designs

PISA uses a rotated booklet design for overcoming the conflicting demands of limited student-level testing time and the broad coverage of the assessment domain. A testing design where students are assigned a subset of items is denoted as an incomplete design. The principles for computing the student's individual ability estimate described in the previous section are also applicable for incomplete designs.

Let us suppose that two students with abilities of –1 and 1 have to answer two out of the four items presented in Table 5.8. The student with $\beta_1 = -1$ has to answer the first two items, *i.e.* the two easiest items and the student with $\beta_2 = 1$ has to answer the last two items, *i.e.* the two most difficult items. Both students succeed on their first item and fail on their second item.

Both patterns have a probability of 0.31 respectively for an ability of –1 and 1. As stated previously, these probabilities can be computed for a large range of student abilities. Figure 5.9 presents the (1, 0) response pattern probabilities for the easy test (solid line) and for the difficult test (dotted line).

Table 5.8

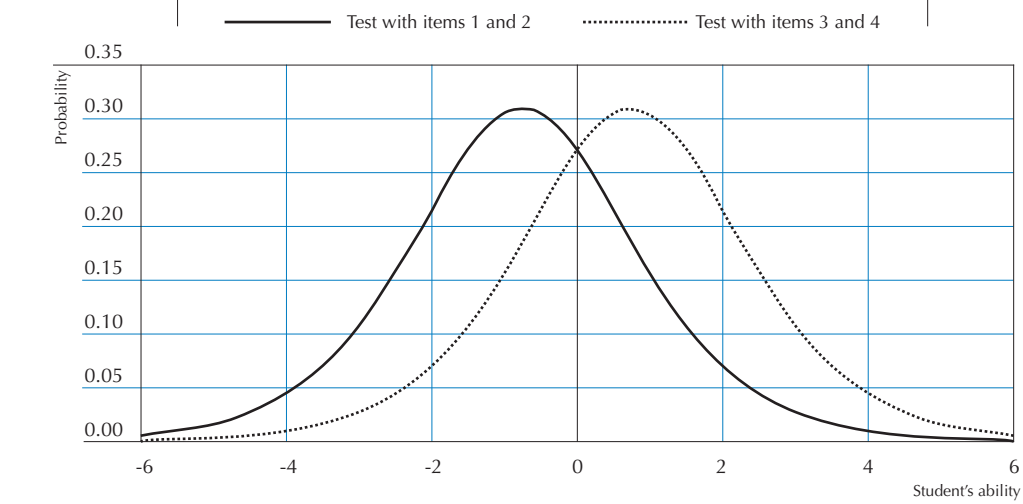
Probability for the response pattern (1, 0) for two students of different ability in an incomplete test design

			$\beta_i = -1$	$\beta_i = 0$
Item 1	$\delta_1 = -1$	Response = 1	0.50	
Item 2	$\delta_2 = -0.5$	Response = 0	0.62	
Item 3	$\delta_3 = 0.5$	Response = 1		0.62
Item 4	$\delta_4 = 1$	Response = 0		0.50
Response pattern			**0.31**	**0.31**

Figure 5.9 shows that for any student that succeeded on one item of the easy test, the model will estimate the student ability at –0.75, and that for any student that succeeded on one item of the difficult test, the model will estimate the student ability at 0.75. If raw scores were used as estimates of student ability, in both cases, we would get 1 out of 2, or 0.5.

Figure 5.9

Response pattern likelihood for an easy test and a difficult test



PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

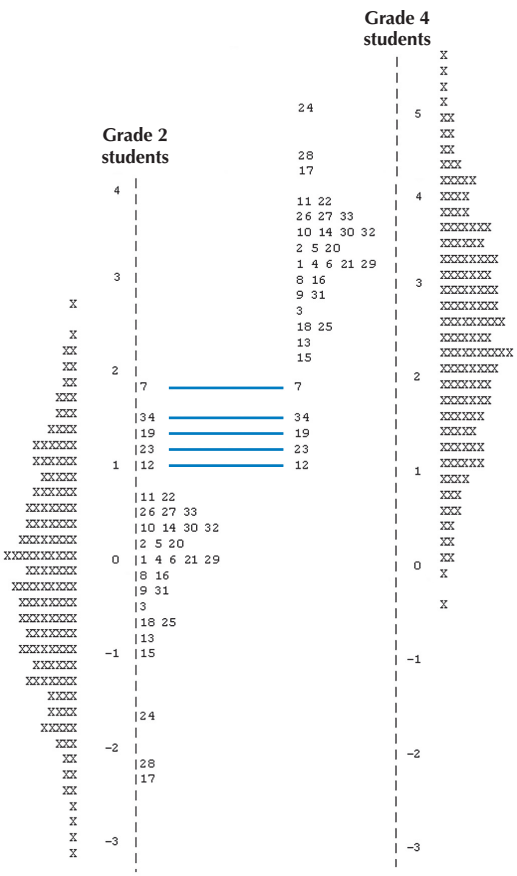
In summary, the raw score does not take into account the difficulty of the item for the estimation of the raw score and therefore, the interpretation of the raw score depends on the item difficulties. On the other hand, the Rasch Model uses the number of correct answers and the difficulties of the items administered to a particular student for his or her ability estimate. Therefore, a Rasch score can be interpreted independently of the item difficulties. As far as all items can be located on the same continuum, the Rasch model can return fully comparable student ability estimates, even if students were assessed with a different subset of items. Note, however, that valid ascertainment of the student's Rasch score depends on knowing the item difficulties.

Optimal conditions for linking items

Some conditions have to be satisfied when different tests are used. First of all, the data collected through these tests must be linked. Without any links, the data collected through two different tests cannot be reported on a single scale. Usually, tests are linked by having different students do common items or having the same students assessed with the different tests.

Let's suppose that a researcher wants to estimate the growth in reading performance between a population of grade 2 students and a population of grade 4 students. Two tests will be developed and both will be targeted at the expected proficiency level of both populations. To ensure that both tests can be scaled on the same continuum, a few difficult items from the grade 2 test will be included in the grade 4 test (let's say items 7, 34, 19, 23 and 12).

Figure 5.10
Rasch item anchoring



PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

Figure 5.10 represents this item-anchoring process. The left part of Figure 5.10 presents the outputs of the scaling of the grade 2 test with items centred on zero. For the scaling of grade 4 data, the reference point will be the grade 2 difficulty of the anchoring items. Then the difficulty of the other grade 4 items will be fixed according to this reference point, as shown on the right side of Figure 5.10.

With this anchoring process, grade 2 and grade 4 item difficulties will be located on a single continuum. Therefore, the grade 2 and grade 4 students' ability estimates will also be located on the same continuum.

To accurately estimate the increase between grades 2 and 4, the researcher needs to ensure that the location of the anchor items is similar in both tests.

From a theoretical point of view, only one item is needed to link two different tests. However, this situation is far from being optimal. A balanced incomplete design presents the best guarantee for reporting the data of different tests on a single scale. This was adopted by PISA 2003 where the item pool was divided into 13 clusters of items. The item allocation to clusters takes into account the expected difficulty of the items and the expected time needed to answer the items. Table 5.9 presents the PISA 2003 test design. Thirteen clusters of items were denoted as C1 to C13 respectively. Thirteen booklets were developed and each of them has four parts, denoted as Block 1 to Block 4. Each booklet consists of four clusters. For instance, Booklet 1 consists of Cluster 1, Cluster 2, Cluster 4 and Cluster 10.

Table 5.9
PISA 2003 test design

	Block 1	Block 2	Block 3	Block 4
Booklet 1	C1	C2	C4	C10
Booklet 2	C2	C3	C5	C11
Booklet 3	C3	C4	C6	C12
Booklet 4	C4	C5	C7	C13
Booklet 5	C5	C6	C8	C1
Booklet 6	C6	C7	C9	C2
Booklet 7	C7	C8	C10	C3
Booklet 8	C8	C9	C11	C4
Booklet 9	C9	C10	C12	C5
Booklet 10	C10	C11	C13	C6
Booklet 11	C11	C12	C1	C7
Booklet 12	C12	C13	C2	C8
Booklet 13	C13	C1	C3	C9

With such design, each cluster appears four times, once in each position. Further, each pair of clusters appears once and only once.

This design should ensure that the link process will not be influenced by the respective location of the link items in the different booklets.

Extension of the Rasch Model

Wright and Masters have generalised the original Rasch Model to polytomous items, usually denoted as the partial credit model (Wright and Masters, 1982). With this model, items can be scored as incorrect, partially correct and correct. The PISA cognitive items were calibrated according to this model.

This polytomous items model can also be applied on Likert scale data. There is of course no correct or incorrect answer for such scales, but the basic principles are the same: the possible answers can be ordered. PISA questionnaire data are scaled with the one-parameter logistic model for polytomous items.

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

OTHER ITEM RESPONSE THEORY MODELS

A classical distinction between item response theory models concerns the number of parameters used to describe items. The Rasch Model is designated as a one-parameter model because item characteristic curves only depend on the item difficulty. In the three-parameter logistic model, the item characteristic curves depend on: *(i)* the item difficulty parameter; *(ii)* the item discrimination parameter; and *(iii)* what can be termed the "guessing" parameter. This last parameter accounts for the fact that, on a multiple choice test, all students have some chance of answering the item correctly, no matter how difficult the item is.

CONCLUSION

The Rasch Model was designed to build a symmetric continuum on which both item difficulty and student ability are located. The item difficulty and the student ability are linked by a logistic function. With this function, it is possible to compute the probability that a student succeeds on an item.

Further, due to this probabilistic link, it is not a requirement to administer the whole item battery to every student. If some link items are guaranteed, the Rasch Model will be able to create a scale on which every item and every student will be located. This last feature of the Rasch Model constitutes one of the major reasons why this model has become fundamental in educational surveys.

Notes

1. See *Measuring Student Knowledge and Skills – A New Framework for Assessment* (OECD, 1999a), *The PISA 2003 Assessment Framework – Mathematics, Reading, Science and Problem Solving Knowledge and Skills* (OECD, 2003b), and *Assessing Scientific, Reading and Mathematical Literacy – A Framework for PISA 2006* (OECD, 2006).

2. The probability of 0.5 was first used by psychophysics theories (Guilford, 1954).

References

Beaton, A.E. (1987), *The NAEP 1983-1984 Technical Report*, Educational Testing Service, Princeton.

Beaton, A.E., *et al.* (1996), *Mathematics Achievement in the Middle School Years, IEA's Third International Mathematics and Science Study*, Boston College, Chestnut Hill, MA.

Bloom, B.S. (1979), *Caractéristiques individuelles et apprentissage scolaire*, Éditions Labor, Brussels.

Bressoux, P. (2008), *Modélisation statistique appliquée aux sciences sociales*, De Boek, Brussels.

Bryk, A.S. and **S.W. Raudenbush** (1992), *Hierarchical Linear Models for Social and Behavioural Research: Applications and Data Analysis Methods*, Sage Publications, Newbury Park, CA.

Buchmann, C. (2000), *Family structure, parental perceptions and child labor in Kenya: What factors determine who is enrolled in school?* a*Soc. Forces*, No. 78, pp. 1349-79.

Cochran, W.G. (1977), *Sampling Techniques*, J. Wiley and Sons, Inc., New York.

Dunn, O.J. (1961), "Multilple Comparisons among Menas", *Journal of the American Statistical Association*, Vol. 56, American Statistical Association, Alexandria, pp. 52-64.

Kish, L. (1995), *Survey Sampling*, J. Wiley and Sons, Inc., New York.

Knighton, T. and **P. Bussière** (2006), "Educational Outcomes at Age 19 Associated with Reading Ability at Age 15", Statistics Canada, Ottawa.

Gonzalez, E. and **A. Kennedy** (2003), *PIRLS 2001 User Guide for the International Database*, Boston College, Chestnut Hill, MA.

Ganzeboom, H.B.G., P.M. De Graaf and **D.J. Treiman** (1992), "A Standard International Socio-economic Index of Occupation Status", *Social Science Research* 21(1), Elsevier Ltd, pp 1-56.

Goldstein, H. (1995), *Multilevel Statistical Models*, 2nd Edition, Edward Arnold, London.

Goldstein, H. (1997), "Methods in School Effectiveness Research", *School Effectiveness and School Improvement* 8, Swets and Zeitlinger, Lisse, Netherlands, pp. 369-395.

Hubin, J.P. (ed.) (2007), *Les indicateurs de l'enseignement*, 2nd Edition, Ministère de la Communauté française, Brussels.

Husen, T. (1967), *International Study of Achievement in Mathematics: A Comparison of Twelve Countries*, Almqvist and Wiksells, Uppsala.

International Labour Organisation (ILO) (1990), *International Standard Classification of Occupations: ISCO-88.* Geneva: International Labour Office.

Lafontaine, D. and **C. Monseur** (forthcoming), "Impact of Test Characteristics on Gender Equity Indicators in the Assessment of Reading Comprehension", *European Educational Research Journal*, Special Issue on PISA and Gender.

Lietz, P. (2006), "A Meta-Analysis of Gender Differences in Reading Achievement at the Secondary Level", *Studies in Educational Evaluation* 32, pp. 317-344.

Monseur, C. and **M. Crahay** (forthcoming), "Composition académique et sociale des établissements, efficacité et inégalités scolaires : une comparaison internationale – Analyse secondaire des données PISA 2006", *Revue française de pédagogie.*

OECD (1998), *Education at a Glance – OECD Indicators*, OECD, Paris.

OECD (1999a), *Measuring Student Knowledge and Skills – A New Framework for Assessment*, OECD, Paris.

OECD (1999b), *Classifying Educational Programmes – Manual for ISCED-97 Implementation in OECD Countries*, OECD, Paris.

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

OECD (2002a), *Programme for International Student Assessment – Manual for the PISA 2000 Database*, OECD, Paris.

OECD (2002b), *Sample Tasks from the PISA 2000 Assessment – Reading, Mathematical and Scientific Literacy,* OECD, Paris.

OECD (2002c), *Programme for International Student Assessment – PISA 2000 Technical Report,* OECD, Paris.

OECD (2002d), *Reading for Change: Performance and Engagement across Countries – Results from PISA 2000,* OECD, Paris.

OECD (2003a), *Literacy Skills for the World of Tomorrow – Further Results from PISA 2000,* OECD, Paris.

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

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

OECD (2004b), *Problem Solving for Tomorrow's World – First Measures of Cross-Curricular Competencies from PISA 2003,* OECD, Paris.

OECD (2005a), *PISA 2003 Technical Report,* OECD, Paris.

OECD (2005b), *PISA 2003 Data Analysis Manual,* OECD, Paris.

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

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

OECD (2009), *PISA 2006 Technical Report,* OECD, Paris.

Peaker, G.F. (1975), *An Empirical Study of Education in Twenty-One Countries: A Technical report. International Studies in Evaluation VIII,* Wiley, New York and Almqvist and Wiksell, Stockholm.

Rust, K.F. and **J.N.K. Rao** (1996), "Variance Estimation for Complex Surveys Using Replication Techniques", *Statistical Methods in Medical Research,* Vol. 5, Hodder Arnold, London, pp. 283-310.

Rutter, M., *et al.* (2004), "Gender Differences in Reading Difficulties: Findings from Four Epidemiology Studies", *Journal of the American Medical Association* 291, pp. 2007-2012.

Schulz, W. (2006), *Measuring the socio-economic background of students and its effect on achievement in PISA 2000 and PISA 2003,* Paper presented at the Annual Meetings of the American Educational Research Association (AERA) in San Francisco, 7-11 April.

Wagemaker, H. (1996), *Are Girls Better Readers. Gender Differences in Reading Literacy in 32 Countries,* IEA, The Hague.

Warm, T.A. (1989), "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika,* Vol. 54(3), Psychometric Society, Williamsburg, VA., pp. 427-450.

Wright, B.D. and **M.H. Stone** (1979), *Best Test Design: Rasch Measurement,* MESA Press, Chicago.

Table of contents

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

LIST OF BOXES

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

LIST OF FIGURES

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

LIST OF TABLES

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009

User's Guide

Preparation of data files

All data files (in text format) and the SPSS® control files are available on the PISA website (*www.pisa.oecd.org*).

SPSS® users

By running the SPSS® control files, the PISA data files are created in the SPSS® format. Before starting analysis in the following chapters, save the PISA 2000 data files in the folder of "c:\pisa2000\data\", the PISA 2003 data files in "c:\pisa2003\data\", and the PISA 2006 data files in "c:\pisa2006\data\".

SPSS® syntax and macros

All syntaxes and macros in this manual can be copied from the PISA website (*www.pisa.oecd.org*). These macros were developed for SPSS 17.0. The 19 SPSS® macros presented in Chapter 17 need to be saved under "c:\pisa\macro\", before staring analysis. Each chapter of the manual contains a complete set of syntaxes, which must be done sequentially, for all of them to run correctly, within the chapter.

Rounding of figures

In the tables and formulas, figures were rounded to a convenient number of decimal places, although calculations were always made with the full number of decimal places.

Country abbreviations used in this manual

AUS	Australia	FRA	France	MEX	Mexico
AUT	Austria	GBR	United Kingdom	NLD	Netherlands
BEL	Belgium	GRC	Greece	NOR	Norway
CAN	Canada	HUN	Hungary	NZL	New Zealand
CHE	Switzerland	IRL	Ireland	POL	Poland
CZE	Czech Republic	ISL	Iceland	PRT	Portugal
DEU	Germany	ITA	Italy	SVK	Slovak Republic
DNK	Denmark	JPN	Japan	SWE	Sweden
ESP	Spain	KOR	Korea	TUR	Turkey
FIN	Finland	LUX	Luxembourg	USA	United States

PISA DATA ANALYSIS MANUAL: SPSS® SECOND EDITION – ISBN 978-92-64-05626-8 – © OECD 2009