Appendix B. Estimation of the environmental science and geoscience performance indices.

The environmental science and geoscience performance indices were constructed using the PISA 2006 dataset employing the Rasch model to estimate students' scores. This appendix describes the process of estimating student performance indices used in the report and outlines differences between the literacy scales used in the initial and thematic PISA reports and performance indices used in this report. In comparison to the method used to produce literacy scales in PISA 2006, a simpler method was employed for this report.

Variables measuring environmental science and geoscience student performance are denoted as performance indices to distinguish them from the original PISA 2006 literacy scales. While both estimates are based on the Item Response Theory (IRT) models they differ significantly in some characteristics, which are discussed below. These differences could potentially affect secondary analysis based on the performance indices which is also discussed in this appendix. Moreover, while PISA science literacy scales are strongly founded in the PISA science literacy framework, such an extensive framework is not present in the case of the environmental science and geoscience performance indices. Thus, the PISA literacy scales and the performance indices developed for this report differ significantly and are not directly comparable. However, they could be used for reliable comparisons of country performance in environmental science and geoscience.

The appendix first discusses the theoretical discrepancies and practical impact of using the simpler way of estimating student performance for this report instead of the original PISA plausible value method. Subsequently, a detailed description of Rasch model estimation is provided. Finally, the construction of proficiency levels and the use of effect sizes in the report are discussed.

The Rasch model for the environmental and geoscience performance indices and the model used to produce PISA literacy scales

The environmental science and geoscience performance indices were constructed using the original PISA 2006 sample with 398750 students from 57 countries. However, it was not possible to calculate the Rasch score for 1266 students (0.3% of the sample) in the environmental science performance index and for 2644 students (0.7% of the sample) in the geoscience performance index. These students didn't provide any answers to items used to construct the performance indices[1]. The plausible value method employed to produce the literacy scales used in the original PISA report imputes missing information assuming a population model. The method used here does not require specifying a population model but also cannot impute missing item information. Thus, students who didn't provide any response on items included in the environmental science or geoscience indices had to be deleted from the analysis. However, the number of missing data is very small and shouldn't affect any of the final conclusions made in the report.

The analysis was conducted using the software WINSTEPS 3.65.0 (www.winsteps.com). This software allows estimating student ability and item difficulty from a pool of items. It also provides estimates of item bias analysis by booklet and it estimates reliability statistics for each level of analysis (student, school, country) which are presented in the following sections. The Rasch model was used to estimate student ability. The Rasch model is discussed in detail in Chapter 5 of the PISA Data Analysis Manuals (OECD, 2009a). It is a basic Item Response Theory one-parameter logistic model where item difficulties and student abilities are estimated simultaneously. The Rasch model is related to the model employed to produce initial estimates used in developing PISA science, mathematics, and reading literacy scales, however, final performance estimates differ as it is outlined below.

In PISA 2006, the method used for the scaling of the student scores was a *"Mixed-coefficients Multinomial Logit Model" (MCMLM)*. This is a categorical response model where the patterns of response to a set of items are modelled as dependent variables through a logistic regression where the predictors are the level of difficulty of the items and the student's ability. This model assumes that the set of unknown parameters is fixed and the student's ability is random. The PISA extension of the model takes into account student background information when modelling latent student ability traits. Furthermore, the model allows drawing a range of plausible values using the population parameters distribution for each scale in PISA. The PISA model estimates several dimensions simultaneously, for example, in PISA 2006 the model to produce literacy scale scores in main domains takes into account responses in science, mathematics, reading, and two attitudinal scales at the same time. This method reduces measurement error and can impute information in case of missing data (OECD, 2009b; Adams & Wu, 2006). Thus, the Rasch model is similar to the MCMLM PISA model because both use the item difficulty and the student's ability to estimate the latent trait behind a set of items. However, they differ importantly in taking into account additional student information about performance on other items and values of background variables (e.g. socio-economic background). The Rasch model does not assume any population distribution, while MCMLM incorporates population model into the estimation of PISA literacy score by using the background student information.

An important difference between the performance indices used in this report and the literacy scales presented in previous PISA reports is that the former assign one ability estimate per student only, while the latter provide five plausible values for each student which are representing a range of abilities that a student might reasonably have. The five plausible values are random draws from the distribution of student ability which takes into account responses to test items and additionally accounts for information provided in the student questionnaire. Analysis with plausible values can limit measurement error in student literacy and provides unbiased estimates of population parameters. In other words, the method of plausible values increases reliability of measurement of student knowledge and skills. Moreover, the plausible values methodology allows imputing literacy scores for students who didn't answer any item on a particular domain. That is not directly possible in the Rasch model used to estimate the environmental science and geoscience performance indices.

Thus, four main differences between the approach used by PISA 2006 and the one used for this report are:

1. The Rasch model was used to estimate item difficulty and student ability instead of the Mixed-coefficients Multinomial Logit Model
2. Only one estimate of student ability is provided instead of drawing five plausible values.
3. The Rasch model used in this report provides student ability estimates based on test item responses only instead of taking into account background student information
4. Missing information about student performance was not imputed.

It is possible that using the Rasch model method instead of the five plausible values literacy scores might affect the analysis undertaken. The following presents some of these points.

Potential impact of different scaling methodology on the secondary analysis of the environmental and geoscience performance indices

The environmental science and geoscience performance indices are based on the maximum likelihood estimates from the one-parameter model. Those have very similar characteristics to warm likelihood estimates (WLE) which were released in the PISA 2000 dataset together with the plausible values estimates. That creates a possibility to compare WLE and plausible values estimates to assess the potential bias in the estimates presented in this report in comparison to using the plausible values analysis present in official PISA reports. While the comparison is based on reading and PISA 2000 it gives an idea about which type of analysis with environmental science and geoscience performance indices is of comparable reliability as the analysis with plausible values estimates of student performance. These issues were already discussed in the technical report and PISA data analysis manuals (OECD, 2009a; OECD, 2009b).

Chapter 6 of the PISA 2006 data analysis manual is devoted to comparison of plausible value (PV) and WLE estimates and discussion below is based on this chapter. The findings presented there were based on a simulation study using PISA data (Monseur and Adams, 2002). These findings could be summarised as follows:

- A good estimate of the population's mean (*i.e.* the latent variable estimate) is obtained regardless of the type of latent variable estimate (WLE or PV).
- The mean of the WLEs will not be biased if the test is well targeted, *i.e.* if the average of the item difficulties is around 0 on the Rasch scale (Wu and Adams, 2002). That is, on a well-targeted test, students will obtain a raw score of about 50% correct answers. If the test is too easy then the mean of the WLEs will be underestimated (this is called the ceiling effect), while if it is too difficult then the mean of the WLEs will be overestimated (this is called the floor effect). These last results explain why the mean of the WLEs provided in the PISA 2000 database differs from the mean of the plausible values, especially for the partner countries. For the reading/reflection and evaluation scale, the means obtained, for example, for Canada using WLEs and PVs are 538.4 and 542.5, respectively, which are very close. In contrast, the means obtained for Peru, using WLEs and PVs are 352.2 and 322.7, respectively, a difference of about 0.3 standard deviations. This shows that there is bias when WLEs are used to estimate the mean if the test is not well targeted.
- For the population variance, PVs give unbiased estimates, while WLEs overestimate it.
- Because the variance computed using plausible values is not biased, the percentiles based on PVs are also unbiased. However, because the WLEs variances are biased, the percentiles, and in particular, extreme percentiles will also be biased.
- It should be noted that the regression coefficients are all unbiased for the different types of estimators. Nevertheless, as variances are biased for some estimators, residual variances will also be biased. Therefore, the standard error on the regression coefficients will be biased in the case of the WLEs and the EAP estimates.
- Between-school variances for the different estimators do not differ from the real values, but WLEs overestimate the within-school variance. Plausible values provide unbiased estimates for both within- and between-school variances.

These findings are helpful in assessing the reliability of the environmental science and geoscience performance indices. Generally, for the countries with mean performance close to the OECD mean, estimates of average performance are only slightly biased. However, for very low performing countries their average student performance could be biased. Nevertheless, the ranking of countries should not change with the more sophisticated scaling approach. Thus, the main message from the report about the position of countries according to their students knowledge and skills related to the environment is not significantly affected by the simpler scaling scheme used for the purposes of this report.

The variance estimates generally will be biased except the between-school variance. This is because of non-negligible error in the measurement of student achievement which diminishes on the aggregate level (e.g. school means). The regression analysis is a reliable method of analysis with these performance indices; however, statistical tests conducted to test hypotheses about the effects of chosen factors could be biased. Thus, the findings should be taken cautiously and compared with other studies. Similarly, analysis for small groups of students with performance levels far from the OECD mean should be also treated cautiously.

Assessing the Rasch model fit for the environmental science and geoscience performance indices

As described above, the Rasch Model focuses on an analysis of the item to estimate the student's ability. For that reason, to assess the adequacy of the scale, it is necessary to evaluate the item fit in the model. In other words, it is necessary to evaluate if correctly answered items with higher difficulty are accomplished by students with higher level of ability. At the same time, these students should have a greater probability of attaining higher scores on items with lower levels of difficulty than on more difficult ones.

Another limitation was the unbalanced number of items by booklet – ranging from four to sixteen according to the booklet. For that reason, an item bias analysis (testing the item difficulty difference) was developed between booklets to see possible problems in the estimation of the parameters for the model. Also, the group reliability by school and country were calculated to indicate the accuracy of the scale to measure the different groups in the dataset.

To assess the adequacy of the items, there are three main statistics in the Rasch Model to evaluate the fit of the item: the items-test correlation (from Classic Test Theory), the mean square statistics (infit and outfit) and the item reliability.

Rasch models use the mean square (MNSQ) fit statistics to identify item and person ratings that deviate from expectations. The MNSQ fit statistics value is the ratio of the observed variance (variance attributable to the observed variable) and the expected variance (variance estimated by the 1PL model). A ratio of 1 indicates that the observed variance equals expected variance; ergo, the error of measurement is almost zero or inexistent. When the MNSQ fit statistics value is greater than 1.0, for example, 1.70, there is 70% more variation in the observed variable than the 1PL model predicted. When the fit statistics value is less than 1.0, there is less variation in the observed variable than the 1PL model predicted.

There are two types of MNSQ fit statistics that have to be taken into account at the moment of analysing the item fit for the scale: the outfit and the infit statistics.

- Outfit mean square is a chi-square statistic that measures the unexpected observations on items that are very easy or very hard for a given individual. In other words, this statistic reflects that there are individuals with higher ability that are answering incorrectly items that should have been easy for them, and vice-versa. It is thought that problems with this indicator do not represent a serious threat to the reliability of the scale (Linacre, 2002).
- Infit mean square is a chi square statistic that measures unexpected patterns of observations by persons on items that are roughly targeted to them. In other words, this statistic evaluates how well the observations fit the IRT model or how large are the residuals in the estimated

model. Problems with this indicator indicate a threat to the reliability and also to the validity of the scale (Linacre, 2002).

These statistics can be categorized as follows according the value of the infit or outfit: off-variable noise is greater than useful information (>2), noticeable off-variable noise (1.5 – 1.99), productive of measurement (0.5 – 1.5), and overly predictable (<0.5). In sum, infit and outfit values between 0.5 and 1.5 indicate that there is a good fit of the items (Linacre, 2008).

As in classic test theory, the item reliability or separation index is the reproducibility of the results or the measure of the item difficulty. Thus, a high item reliability index means that there is almost no error in the estimation of the item difficulty and this will be very close to the observed one. Values close to 1 mean a reliable measure and values close to 0 mean that the measurement error is greater than the variance explained by the model.

Psychometric characteristics of the items in the environmental sciences scale

As previously mentioned, the item fit was evaluated in function of three indicators: item-test correlation, mean square fit statistics (outfit and infit) and item reliability or separation index. The criteria for each indicator were:

▪ Item-Test correlation:	above 0.20
▪ Infit-statistic:	between 0.50 and 1.50
▪ Outfit-statistic:	between 0.50 and 1.50

An item which presented any problem with these indicators was discarded from the scale. The analysis developed shows that all the items considered for the environmental science scale and the geosciences scale had a good fit and it was not necessary to delete any item from the scale. Then, the overall fit of the items was evaluated through the item separation index, and in both scales the value of this indicator was equal to 1. This means that there is almost no error in the estimation of the item difficulty.

Also, given the fewer number of item to assess the environmental and geosciences knowledge of the students, in addition to a person reliability, a group reliability was also estimated (Linacre, 2008). The group reliability gives an indication of the accuracy of the scale to compare different groups, in this case schools and countries. In the case of the group reliability (by school or country), it was above 0.80 indicating a satisfactory measure at these levels; while at student level, reliability index was lower than 0.30. Thus, the analysis developed shows that there is low person reliability for both scales but higher group reliability. This is a common result in international evaluation where the main objective is to compare the proficiency of students grouped by countries.

Table 1. Item fit statistics for the environmental science scale (descending order)

	# item	**Logit score**	**Infit mnsq**	**Outfit mnsq**	**Item-test correlation**
gs114q05t	3	1.76	1.08	1.10	0.44
es458q01	17	1.71	1.03	1.05	0.44
es425q04	14	1.18	0.99	0.95	0.54
es268q02t	4	0.71	0.94	0.90	0.59
es269q04t	8	0.69	1.08	1.21	0.46
gs114q04t	2	0.68	0.77	0.69	0.74
gs304q03a	9	0.42	0.95	0.91	0.56
gs465q04	20	0.42	1.16	1.27	0.45
es269q03t	7	0.31	0.86	0.78	0.60
es425q03	13	0.30	1.04	1.04	0.54
es425q02	12	0.20	1.03	1.08	0.55
gs514q03	23	0.09	0.96	0.94	0.59
es408q04t	10	-0.12	1.10	1.13	0.51
gs465q01	19	-0.20	0.89	0.84	0.72
gs114q03t	1	-0.32	0.84	0.79	0.67
gs527q04t	24	-0.35	1.14	1.21	0.46
es268q06	5	-0.38	1.05	1.06	0.56
es458q02t	18	-0.52	1.05	1.08	0.54
gs466q05	21	-0.55	1.17	1.23	0.47
gs485q02	22	-0.60	0.97	0.95	0.56
gs269q01	6	-0.74	0.88	0.84	0.60
gs426q03	15	-1.17	1.10	1.12	0.49
gs426q05	16	-1.65	0.91	0.84	0.57
gs415q02	11	-1.87	0.99	0.99	0.55

Item reliability:	1.00
Person reliability (Student):	0.33
Group reliability (School):	0.89
Group reliability (Country):	0.99

Table 2. Item fit statistics for the geosciences scale (descending order)

	# item	**Logit score**	**Infit mnsq**	**Outfit mnsq**	**Item-test correlation**
gs114q05t	3	2.24	1.13	1.19	0.48
gs114q04t	2	1.09	0.73	0.68	0.78
gs304q03a	9	0.84	1.00	0.99	0.61
gs465q04	20	0.71	1.19	1.29	0.48
gs514q03	23	0.46	0.99	0.98	0.69
gs465q01	19	0.06	0.80	0.76	0.77
gs114q03t	1	-0.05	0.86	0.81	0.70
gs527q04t	24	-0.08	1.14	1.21	0.51
gs466q05	21	-0.28	1.18	1.23	0.55
gs485q02	22	-0.34	0.98	0.98	0.59
gs269q01	6	-0.50	0.91	0.88	0.62
gs426q03	15	-0.95	1.11	1.15	0.52
gs426q05	16	-1.46	0.92	0.86	0.60
gs415q02	11	-1.73	1.03	1.12	0.57

Item reliability:	1.00
Person reliability (Student):	0.15
Group reliability (School):	0.85
Group reliability (Country):	0.98

Once the model fit of the items had been established, it was possible to evaluate bias in the item difficulty by booklet. As in the majority of international evaluations, PISA uses a system of rotated booklets; in other words, not all the students answer all the items because each item appears four times across booklets (thirteen in total). This method estimates the score of the students only with the information from the items that they answer.

However, it is important to ensure that there is no item bias across booklets; in other words, the item difficulty does not vary across the booklets where the item is present. The cut-off score to consider items being biased between booklets was when the difference in difficulty was +/- 0.50 logits (Wright & Douglas, 1976; Wright & Stone, 1979). Then, the item difficulty was estimated in each booklet (according the design of the test, each item appears in four booklets) and an item in a booklet was eliminated when: a) its item difficulty differs with more than two booklets, and b) in the case of an item that is different in two booklets, the item is deleted from the booklet which had the biggest difference with the average item difficulty. During the analysis, it was found that there were only three cases where item difficulty bias was observed across booklets and the item was deleted from the booklet with the biggest difference. Thus, the deletion of these items in the different booklets does not affect the item fit (infit, outfit and point-biserial correlation) or the order of the items in the test (ranking).

Table 3: Item bias analysis by booklet for the environmental science scale

		Booklet												
# item	Label	1	2	3	4	5	6	7	8	9	10	11	12	13
1	gs114q03t	-0.29	-0.39						-0.40			-0.04		
2	gs114q04t	0.60	0.72						0.68			0.93		
3	gs114q05t	1.74	1.86						1.69			1.97		
4	es268q02t		0.70	0.57		0.76				0.76				
5	es268q06		-0.40	-0.46		-0.37				-0.40				
6	gs269q01	-0.71								-0.67	-0.67		-0.67	
7	es269q03t	0.38								0.38	0.40		0.33	
8	es269q04t	0.70								0.77	0.87		0.74	
9	gs304q03a					0.56	0.16				0.28			
10	es408q04t	-0.06		-0.05	-0.39		-0.09							
11	gs415q02			-2.23	-2.48									
12	es425q02	0.37				0.15		0.28						0.08
13	es425q03	0.46				0.20		0.27						0.37
14	es425q04	1.23				1.03		1.27						1.30
15	gs426q03	-1.10								-1.10	-1.06		-1.15	
16	gs426q05	-1.65								-1.62	-1.47		-1.63	
17	es458q01					1.52	1.90		1.99		1.62			
18	es458q02t					-0.50	-0.40		-0.45		-0.63			
19	gs465q01				-0.21	-0.27						-0.55		
20	gs465q04				0.45	0.41						0.18	0.59	
21	gs466q05					-0.47	-0.60		-0.45		-0.57			
22	gs485q02	-0.61								-0.56	-0.43		-0.53	
23	gs514q03	0.26				0.07		0.02						0.11
24	gs527q04t	-0.28								-0.28	-0.28		-0.28	

Table 4. Item bias analysis by booklet for the geosciences scale

		Booklets												
# item	Label	1	2	3	4	5	6	7	8	9	10	11	12	13
1	gs114q03t	-0.08	-0.41						-0.20					
2	gs114q04t	0.94	1.08						1.01			1.33		
3	gs114q05t	2.14	2.37						2.08			2.35		
6	gs269q01	-0.51								-0.54	-0.54		-0.60	
9	gs304q03a					0.91	0.82		1.17					
11	gs415q02	-1.39		-1.39			-1.39							
15	gs426q03	-0.94								-1.00	-0.97		-1.09	
16	gs426q05	-1.51								-1.55	-1.40		-1.57	
19	gs465q01				-0.03	0.01						-0.18	0.22	
20	gs465q04				0.78	0.73						0.56	0.65	
21	gs466q05					-0.22	-0.15		-0.24		-0.45			
22	gs485q02	-0.41								-0.39	-0.29		-0.46	
23	gs514q03	0.54				0.38		0.45						0.45
24	gs527q04t	-0.04								-0.07	-0.14		-0.23	

Finally, once the item bias by booklet in each scale was corrected, the Rasch scores were estimated for each student. Then, these Rasch scores were rescaled using senate weights (giving equal weight to each country) and data from OECD countries to make the OECD mean 500 and the standard deviation across the OECD countries being 100.

The development of new proficiency levels for the environmental science and geoscience performance indices.

1. For the purpose of describing what students know and can do in terms of environmental science and geoscience, proficiency levels were also developed.

2. Whenever possible, the design of these proficiency levels followed the techniques used to develop the proficiency levels for science in PISA 2006 (PISA 2006 Technical Report, OECD, 2009b).

3. There are, however, two important differences between the PISA 2006 science proficiency levels and the proficiency levels described in this report. The environmental science and geoscience performance indices used fewer test items than the overall PISA 2006 science scale. In addition, the test items used to develop the new indices did not span the whole range of proficiency levels used in reporting science performance in PISA 2006. For example, there were no environment related questions in the lowest level of science proficiency (Level 1). The process resulted in the same four levels of proficiency for both environmental science and geoscience. To distinguish them from the proficiency levels in PISA 2006, this report refers to proficiency Levels A (the highest level) to D (the lowest level).

4. A three-step procedure was used to develop the proficiency levels. First, the 24 questions were ordered from the least to the most difficult to answer. The relative difficulty of questions in a test is estimated by considering the proportion of test takers getting each question correct. The result is a set of estimates that allows the creation of a continuum representing environmental science and geoscience competencies. On this continuum it is possible to place individual students, thereby showing the degree of environmental science and geoscience proficiency they demonstrate. It is also possible to place individual questions, thereby showing the level of science proficiency each questions embodies.[2]

5. Second, a group of environmental science and geoscience experts evaluated the ranked list of questions, grouping the questions into proficiency levels on the basis of the underlying abilities required to answer the questions. The experts used three criteria based on the PISA 2006 science framework as applied to environmental science and geoscience:

- The degree to which the item required students to apply their own knowledge of environmental and/or geoscience to answer the item correctly
- The degree to which the item required the student to understand the interaction among components of an environmental ecosystem to answer the item correctly
- The degree to which the item required the student to synthesise and use a problem-solving strategy applied to a specific environmental or geoscience issue in order to answer the item correctly

6. These item evaluations were used to form four levels of student proficiency in environmental science and geoscience, from Level A being the most proficient to Level D being the least proficient.

7. Third, the experts compared how the 24 environmental science and geoscience questions were distributed on the four proposed proficiency levels as against how they were distributed on the six PISA 2006 science proficiency levels. The comparison showed considerable agreement.

Cut-off scores for the environmental science scale.

8. To determine the cut-off scores for the environmental science scale, the following steps were followed: a) the items were ordered from the easiest to the hardest according their Rasch scores or item difficulty, b) the items were reviewed and the proficiency levels determined according the difficulty of the items and the requirements of each task, c) the different proficiency levels were calculated such that at the bottom of the scale a student at that level would have a 50% chance of getting the items correct at that level.

9. Finally, in the case of the geo-science scale, the same cut-off scores were used for the calculation of the proficiency levels.

10. The following chart illustrates the proficiency levels and the score necessary to be in each level.

Table 5. Item difficulty, Proficiency levels and cut-off scores for the environmental science and geosciences performance indices

Item Label	Item difficulty 50% of get right the item	Performance levels		Cut-off scores with 62% of get right the item
		Environment scale	Science scale	
gs415q02	354.9	1	2	>386.9 (LEVEL D)
gs426q05	404.6	1	2	
gs426q03	436.6	1	2	468.6
gs269q01	464.7	2	3	>468.6 (LEVEL C)
gs485q02	473.8	2	3	
gs466q05	474.5	2	3	
es458q02t	475.8	2	3	
es268q06	482.3	2	3	
gs465q01	486.9	2	3	
gs114q03t	489.5	2	3	
gs527q04t	490.1	2	3	
es408q04t	498.6	2	3	530.6
gs514q03	515.6	3	4	>530.6 (LEVEL B)
es425q02	522.8	3	4	
es425q03	529.3	3	4	
gs304q03a	530.6	3	4	
es269q03t	533.3	3	4	
gs465q04	535.2	3	4	
es268q02t	554.2	3	4	
gs114q04t	555.5	3	4	
es269q04t	558.7	3	4	590.7
es425q04	586.8	4	5	>590.7 (LEVEL A)
es458q01	622.8	4	6	
gs114q05t	626.0	4	6	

Effect sizes

Sometimes it is useful to compare differences in an index between groups, such as males and females, across countries. A problem that may occur in such instances is that the distribution of the index varies across countries. One way to resolve this is to calculate an effect size that accounts for differences in the distributions. An effect size measures the difference between, say, the awareness of environmental issues of male and female students in a given country, relative to the average variation in student awareness among male and female students in the country.

An effect size also allows a comparison of differences across measures that differ in their metric. For example, it is possible to compare effect sizes between the PISA indices and the PISA test scores, as when, for example, gender differences in performance in science are compared with the gender differences in several of the indices.

In accordance with common practices, effect sizes less than 0.20 are considered small in this volume, effect sizes in the order of 0.50 are considered medium, and effect sizes greater than 0.80 are considered large. Many comparisons in this report consider differences only if the effect sizes are equal to or greater than 0.20, even if smaller differences are still statistically significant; figures in bold in data tables presented in this report indicate values equal to or greater than 0.20. Values smaller than 0.20 but that due to rounding are shown as 0.20 in tables and figures have not been highlighted. Light shading represents the absolute value of effect size is equal or more than 0.2 and less than 0.5; medium shading represents the absolute value of effect size is equal or more than 0.5 and less than 0.8; and dark shading represents the absolute value of effect size is equal or more than 0.8.

The effect size between two subgroups is calculated as:

$$\frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$$

m_1 and m_2 respectively represent the mean values for the subgroups 1 and 2. σ_1^2 and σ_2^2 respectively represent the values of variance for the subgroups 1 and 2. The effect size between the two subgroups 1 and 2 is calculated as dividing the mean difference between the two subgroups (m_1 - m_2), by the square root of the sum of the subgroup's variance ($\sigma_1^2 + \sigma_2^2$) divided by 2.

References

Adams, R.J., & Wu, M.L. (2006). *The mixed-coefficient multinomial logit model: A generalized form of the Rasch model.* In M. von Davier & C.H. Carstensen (Eds.) Multivariate and mixture distribution Rasch models: Extensions and applications, (pp. 57 – 76). Springer Verlag.

Baker, F. (2004). *Item response theory: parameter estimation techniques (2nd. ed.)* New York : M. Dekker.

Linacre, J. M. (2002). *What do infit and outfit, mean-square and standardized mean?* Rasch Measurement Transactions, 16(2), 878.

Linacre, Mike (2008). A user's guide to WINSTEPS MINISTEPS Rasch Model Computer Programs. Chicago: MESA Press.

OECD (2009a). PISA Data Analysis Manuals. OECD, Paris.

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

Wright, B.D. & Stone, M. (1979). *Best Test Design*. Chicago: MESA Press.

Wright, B.D. & Douglas, G.A. (1976). *Rasch item analysis by hand. Research Memorandum Number 21*, The University of Chicago, Department of Education, Statistical Laboratory.

[1] The number of items used to construct environmental science and geoscience performance indices was relatively small and these items were unevenly spread over several booklets (see tables 3 and 4 in this Appendix). It is possible that students may not have answered any of the environmental science and geoscience items. For these students, no estimates of performance are available.. Therefore, they were not included in the analysis. Their omission, however, does not significantly affect results as this group of students constitute less than 1% of the OECD student population. On average in OECD countries 0.55% of students were not included in the analysis. In a few countries the number of students missing was slightly higher, reaching 3% in Czech Republic, Germany, and Slovenia. To check whether that could affect final country scores, the original PISA science literacy scale score was calculated for a subsample of students with non-missing values of the environmental science performance index and compared with a score for all students sampled in PISA. That exercise gave an indication of the impact that excluding students with missing item responses could have on country ranking. It was found that the ranking of countries remained almost the same - only two changed their ranking by more than 1 position, while maintaining practically the same mean score. Thus, it is reasonable to assume that the exclusion of students with missing responses did not affect country scores significantly and has a negligible impact on the analysis.

[2] This does not mean that students will always be able to perform questions at or below the difficulty level associated with their own position on the index, and never be able to do harder questions. Rather, the ratings are based on probability: A student with a given score on the index is likely to get a question with the same score correct.