Dealing With Differential Item Functioning In Item Response Models: Inter-temporal Differences

James McIntosh

Economics Department
Concordia University
1455 De Maisonneuve Blvd. W.
Montreal, H3G 1M8,
Quebec, Canada.
james.mcintosh@concordia.ca.

and

623 Econometrics,
601 Mian Rd.
Hudson J0P 1J0
Quebec, Canada
623econometrics@gmail.com.

Abstract

This paper examines the problems which arise in the evaluation of academic performance indicators over time. A two parameter model is applied to PISA mathematics item data for Canada for the years 2009 and 2012. In the estimation procedure item difficulty and dispersion parameters were allowed to differ across the two years. Different normalizations for identifying the distribution parameters were also considered. The choice of normalization is crucial in guaranteeing certain invariance properties required by item response models. The presence of inter-temporal DIF requires the estimation of regression and distribution parameters and the computation of ability scores simultaneously for both years. The ability scores obtained from the methods employed here are significantly higher in 2012 in sharp contrast to PISA results, which gave lower scores for 2012.

Introduction

The Programme For International Student Assessment (PISA) is an OECD (2010) sponsored initiative to allow member and associated countries to evaluate their educational systems by administering a common set of proficiency tests in science, mathematics, and reading to samples of respondents aged 15. The survey results make it possible for countries to rate the performance of their educational system and how this changes over time. It also allows them to assess the importance of key variables, like those representing the respondent's social and economic background, school characteristics, and geographical or administrative region, in the determination of academic performance[1].

The impact of PISA has been enormous. Concerns about the success of educational policy have risen dramatically in participating countries, especially in most of Western Europe, where performance is substantially below Finland and the leading Asian countries. PISA results have shocked both educationalists and politicians whose views about the quality of education were much more optimistic and favourable than those conveyed by PISA.

Researchers in this area have focused exclusively on dealing with problems that arise when two distinct groups like different sub-populations or in the case of PISA where different countries are being compared. But countries, in particular, are concerned whether their academic performance in getting better or worse over time. Canadian PISA results showed a significant downward trend in scores. PISA scores in mathematics for Canada were 527 in 2009 and 518 in 2012 and Canada's rank in mathematics dropped from 9th to 13th place over this three year period tied with Finland OECD (2014: 18). As has been the case in many other countries these results have been at the forefront of education policy discussion. Many Canadian educators have expressed similar concerns about what they see as a serious decline in the quality of mathematics skills of Canadian students. Declines in some of the provincial scores were even larger generating more distress.

Are these concerns justified? There is a growing list of PISA doubters. Many other researchers have expressed reservations about PISA methodology. A short list includes Kreiner and Christensen (2013): unresolved DIF problems and concerns about the way plausible values are calculated, Prais (2007): sampling problems, and Goldstein et al (2007): absence of multivariate random effects and improper treatment of school effects. Wuttke (2007, p. 1) is particularly negative towards PISA: biased samples, item deviation from the Rasch model, non-uniform coding etc.

[1] PISA data is publically available to anyone who wants to use it and can be found at the websites: pisa2009.acer.edu.au and pisa2012.acer.edu.au.

This paper is also critical of the way PISA computes ability scores. However, the problems that are associated with PISA methodology arise in all item response models when certain basic procedures are not followed. It was shown in McIntosh (2017a) that the type of normalization used to identify the distribution parameters is crucial in preserving the fundamental invariance features of contemporary item response models[2]. PISA uses an item free normalization (in PISA the difficulty parameters sum to zero, OECD (2010: 229)) and this means that ability scores depend on the actual values of the difficulty parameters. This is not a good normalization to use and one proposed by Verhelst and Glas (1991) should be used instead. As a result differences in difficulty parameters between countries or between two time periods will contribute to spurious score differences. This is not a problem which is unique to PISA but one which one which affects all item response models whose distribution parameters are improperly normalized.

This research examines the claim based on the scores which PISA publishes every three years that Canadian mathematical skills are in decline. Using methods which deal with some of the problematic aspects of PISA methodology it is shown that, contrary to published PISA scores, Canadian mathematics scores have actually improved over this period. The main result that emerges from reworking the PISA data is that Canadian mathematics test score performance is significantly better in 2012 than it was in 2009 and Canadians should be celebrating its success in mathematics as one of PISA's better performers.

To briefly summarize the results obtained here, the random effects Rasch (1960) model which is the analytical tool used by PISA in the production of ability scores relies on assumptions which are not satisfied by the data which is used to estimate the parameters in the model. See OECD (2009: Ch. 9) for a detailed description of the PISA model. Rasch models are always rejected in favour of the more general two parameter logistic (2PL) models. Parameter estimates and ability scores are dependent on the normalization used to identify the distributional parameters. Comparing results over two time periods means that a normalization has to be found which minimizes the effect of differences in the difficulty parameters over the two periods. This is a radical improvement in what PISA does since the method recognizes the need to estimate the two sets of parameters and scores simultaneously over the two periods.

IRT Methodology

PISA tests consist of a series of questions or items. All major testing programmes use the item data rather than an aggregate score like the raw item score which is just the sum of the correct answers on the group of items under consideration. In PISA most of the item answers are binary; that is to say that item answers are considered as being either right or wrong.

[2] These are the independence of ability scores and distribution parameters and the invariance of the score with respect to the choice of items; changing an item in the test should not change the score.

An IRT model is based on an error components latent variable construction. Suppose that there are J items under consideration and V individuals answering the test. Define the latent variable

$$y_{vj}^{*} = \theta_v + u_{vj} - \delta_j \quad (1)$$

Where Θ_v is individual v's ability. Ability has a deterministic and a random component making

$$\theta_v = \mu_v + \varepsilon_v \quad (2)$$

u_{vj} and ε_v are random components and δ_j is the difficulty level of item j. Θ_v - δ_j can be considered as a measure of net ability; when net ability is positive correct answers are the outcome. Let the cumulative distribution of u_{vj} be F_j. Let y_{vj} be the response to item j. $y_{vj} = 1$ if v gets a correct answer for item j otherwise $y_{vj} = 0$. Conditional on ε_v, $y_{vj} = 0$ if $y_{vj}^{*} < 0$ otherwise $y_{vj} = 1$. The outcome probabilities, therefore, become

$$\Pr\{y_{vj} = 0\} = F_j(\delta_j - \theta_v) = P_{vj} \quad (3)$$

$$\Pr\{y_{vj} = 1\} = 1 - F_j(\delta_j - \theta_v) = 1 - P_{vj} \quad (4)$$

Assuming that, conditional on ε_v, the item outcomes are independent then the conditional likelihood function is

$$L(\alpha,\beta,\delta \mid \varepsilon) = \prod_{v=1}^{V} \prod_{j=1}^{J} P_{vj}^{(1-y_{vj})}(1-P_{vj})^{y_{vj}} \quad (5)$$

Since ε is not observable the average or integrated likelihood function

$$L(\alpha,\beta,\delta) = \prod_{v=1}^{V} \int \prod_{j}^{J} P_{vj}^{(1-y_{vj})}(1-P_{vj})^{y_{vj}} \phi(\varepsilon_v) d\varepsilon_v$$

(6)

is required for estimating the model parameters. Φ( ) is the mixing distribution which is often assumed to be univariate normal with variance parameter σ.

The assumptions of conditional independence of the items and that ability is the same for all items and item difficulty is the same for all individuals are common in IRT models and have their origin in the work of Rasch (1960) and Lord and Novick (1968). The use of two random effects may appear strange to some readers but there are compelling reasons for following this procedure. Items are not independent. Getting item j correct makes more likely that other items will be answered correctly. But joint distributions of correlated binary random variables are very uncommon and using continuous multivariate distributions to model the outcome categories becomes increasingly complex as the number of items increases. Even with modern simulation methods the computational burden makes it impractical for use on this type of data when there are many items.

The correlation between the items is achieved by employing a mixing procedure. The random effect ε_v which is common to all of the item probabilities for individual v is integrated out of the likelihood function using numerical methods described in Aitken and Aitken (2011) or Heckman and Singer (1986) when there is there is some doubt over the form of the mixing distribution, $\varphi(\varepsilon_v)$. For PISA ε_v is univariate but other large tests use multivariate random effects. This procedure is efficient and relatively simple to implement.

IRT and PISA

To do actual empirical work functional forms and algebraic representations have to be determined. PISA uses a logistic distribution to represent the item probabilities.

$$F_j() = \frac{\exp(\alpha_j(\theta v - \delta j))}{1 + \exp(\alpha_j(\theta v - \delta j))} \quad (7)$$

This is called the two parameter logistic distribution. PISA uses a one parameter version of it where all of the α_j parameters are set equal to unity. The α_j are called dispersion parameters and measure the variances of the item outcomes. The deterministic part of ability is assumed to depend on the individual's observable characteristics, x_{vk}. The IRT model of PISA is thus similar to the regression model with

$$\mu_v = \beta_0 + \sum_{k=1}^{K} x_{vk}\beta_k = X_v\beta \quad (8)$$

Like the regression model the β coefficients are the same for all individuals.

The model produces a score which can be used to represent θ_v. This is due to Lindsay et al (1995) and is the Bayesian conditional of θ_v given y_v or

$$\theta_v^* = \mu_v + E(\varepsilon_v \mid y_v) \quad (9)$$

All probability models require a normalization to identify the distributional parameters. PISA uses the normalization

$$\sum_{j=1}^{J} \delta_j = 0 \quad (10)$$

to identify the difficulty parameters. All the dispersion parameters are equal to unity so there are no identification issues with them. This is where PISA's problem begin.

As was shown in McIntosh (2107a) under the normalization described by equation (10) the θ^* scores depend on the difficulty of the items. In these IRT models individual ability and item difficulty are quite separate constructs and deleting or changing an item from the test should not alter the score or have any effect on the estimates of the other difficulty parameters. This does happen under this normalization and so an alternative normalization needs to be found. Verhelst and Glas (1995) proposed

$$(\alpha_j, \delta_j) = (1,1) \quad (11)$$

for some j. When this is used the θ^* scores do not depend on which items appear in the test. Only Verhelst-Glas normalizations produce parameter estimates and scores which are invariant to the level of item difficulty. This result ensures that in the comparison of test results for two different years the scores will not be affected by the possibly different levels of item difficulty even when the same items are used in each year.

The parameters that are actually estimated when this normalization is used are similar to those for any other probability model: the estimated regression parameters are of the form $\alpha_j\beta$, and the distribution parameters are of the form α/α_j and δ/δ_j. This means that regression parameter estimates and Θ^* scores depend on the normalization used. It is also the case that the normalization does not affect the maximized value of the likelihood function. How j is chosen will be discussed below.

PISA Item Data

The way that PISA elicits information from the students who participated in the programme presents some challenges for its analysis. Students answer a set of items which are contained in a booklet. Each booklet, of which there are thirteen, contains four clusters. A cluster is a group of items all of which belong to the same subject. Booklets were randomly assigned and the average number of students covered in each booklet was about 1785 in 2009 and 1655 in 2012. There are three subjects: reading, mathematics, and science. PISA 2009 focused on reading and there were only three mathematics clusters. PISA 2012 focused on mathematics and there are seven mathematics clusters for that year. Clusters 1, 2, and 3 had the same items in each year and the analysis will be restricted to those three clusters keeping in mind that the respondents were not the same over the two years. They were aged between 15.3 and 16.6 years. Only students in grade 9 or 10 are included.

Carrying out the statistical analysis on booklets is not a good research strategy. Considering several subjects in an item response model creates a potential inefficiency problem if ability is not the same in each subject, a condition that could reasonably be expected to characterize the students in the PISA sample. When ability is subject specific then the model has to have an ability function and a set difficulty parameters for each subject in the booklet. Given that booklet sample sizes are quite small this is not a reliable procedure since there are not enough observations to estimate all of the parameters efficiently.

Estimating the model using clusters is a much better alternative. Average sample sizes for clusters are 7140 and 6620 for 2009 and 2012, respectively. These are much larger than booklet sample sizes and are more suitable for the estimation of the parameters in the item response models discussed above. But this approach is not without its difficulties. Respondents who answered the items in mathematics 2, for example, do not form a homogeneous group because they answered items from different booklets and the booklets were not the same in each year. Booklets can vary in difficulty and have different alternatives to mathematics, some with both reading and science others with just reading or just science. In all of the versions of the models which contain regressors three booklet identifiers are included as covariates. This is not an ideal solution but appears to be the best available. In any case, booklet effects are always highly significant indicating that more control for the presence of other subjects in the booklet is required.

Regressors are always included. The variables representing the respondents' characteristics are father's and mother's education, the socioeconomic status of the respondent's family, age, grade, and gender.

Estimation Procedures

In order to estimate the model a decision has to be made concerning the choice of the item to be used in normalizing the distribution parameters. Let (A, B) represent the years (2009, 2012). It was shown in McIntosh (2017a) that if there is an item which satisfies the condition $(\alpha_j^A, \delta_j^A) = (\alpha_j^B, \delta_j^B)$ for some j in {1,2..,J} then imposing the Verhelst-Glas rule on this item will give the correct scores and parameter estimates for both years, A and B, even if all of the other distribution parameters differ over the two years. It is also the case that the maximized likelihood function is higher for this item than for any other item. To find the appropriate item the rule was applied to all of the items and the item with the largest maximized value of the likelihood function was chosen. This is called the anchor item, item a. For item a $(\alpha_a , \delta_a) = (1,1)$ for both years. Of course, there may be no such item for which the distribution parameters are equal in which case all that can be said is that the anchor item is the one for which the distribution parameters are most similar. This item choice minimizes the effect of differences in distribution parameters across the two year but it may not be free of DIF. However, as is shown below, it is possible to determine whether the anchor item is free of DIF.

Unlike PISA the two sets of parameters are estimated for both years simultaneously and the scores are also computed in this manner. In PISA each year is treated separately and nothing is done to deal with the inter-year differences in the distribution parameters.

For cluster 1 the anchor item was M442Q02. The difficulty parameter estimates shown in Table 1 come from maximizing the likelihood function using the pooled data for the two years. Difficulty, dispersion, and regression parameters were allowed to be different across the two years. The item success rates are similar for the two years and average difficulty parameters are as well, although there are significant individual item differences in the difficulty parameters. Unlike PISA the dispersion parameters were not constrained to be equal to unity. The averages for the two years are 1.24 (0.03) and 1.23 (0.02), respectively and are significantly different from unity as are most of the individual dispersion parameter estimates.

It is interesting to note that the PISA normalization rule requires that average difficulty has to be the same for each year. In addition to causing problems associated with ability being dependent on item difficulty it is a restriction which is not satisfied by the data in cluster 2. Biases in the estimated parameters will be the result of imposing this invalid restriction.

For the other two clusters the results were similar. There no significant differences in average difficulty although there were some significant differences for individual items.

As noted earlier it is important to determine whether the anchor item is free of DIF. Swamanathan and Rogers (1999) proposed a simple single item procedure to determine whether there were group differences in the item's difficulty. They suggested applying the logistic probability model to the tem in question. Their equation for item a is

$$\Pr\{y_{va}=1\}=1/(1+\exp(\gamma_0+\gamma_1 S_v+\gamma_2 g_v+\gamma_3 S_v g_v) \quad (12)$$

In (12)

$$S_v=\sum_{j=1}^{J} y_{vj} \quad (13)$$

is the raw item score and g_v is the year indicator. In McIntosh (2018) it was shown that this procedure produces biased results because of the correlation between the raw item score and the unobservable error term in equation (2). As an alternative an amended Swamanathan–Rogers test was proposed where S_v is replaced by θ_v^* in equation (12) and the model is estimated by non-linear least squares. Equation (12) is replaced by

$$y_{va}=1/(1+\exp(\gamma_0+\gamma_1 S_v+\gamma_2 g_v+\gamma_3 S_v g_v)+\eta_{va} \quad (14)$$

Where η_{va} is a random error term.

θ_v^* still depends on ε_v but an instrumental variable procedure can be used to deal with that. The results of applying this test to the three mathematics clusters are shown in Table 3. The instruments are the regressors used in the computation of θ_v^*.

PISA Results

Table 2 gives the results for three different representations of mathematical ability. In this table all scores are weighted by the student weight variable. The first is what PISA determined ability should be. These are called plausible value scores and they are drawings from the same conditional distribution that generate the aggregate country scores and rankings. Here only plausible value scores for respondents who actually answered the questions in the respective cluster were included. The second is the raw item score defined in equation (12) which is just the sum of the correct answers. The third score is a PISA like renormalized version of θ_v^*.

$$P_v^* = 500 + 100[\theta_v^* - Mean(\theta^*)] / \sqrt{Var(\theta^*)} \quad (13)$$

As Table 3 shows, the amended Swamanathan-Rogers tests confirm that all three anchor items are free of DIF so there is no evidence that the distribution parameters differ across years for these three items.

When PISA plausible value scores are restricted to those who answered items in mathematics all three clusters show significantly higher scores for 2012. The same is true for the other two scores. They are all significantly higher in 2012 than in 2009. For all three clusters average difficulty was about the same for the two years.

Thus, there is no evidence that Canadian mathematical ability has declined over this three year period. In fact, contrary to PISA's results there was a significant increase in ability. Based on the P^* score Canada's 2012 PISA score should be revised upward to 537 which places it between Japan, seventh and Macau, sixth as one of the PISA top performers. Alternatively, based on the actual PISA scores in the third column of Table 2 Canada's PISA score only rises to 533 which places it between Switzerland, ninth and Lichtenstein, eighth. Of course, some caution needs to be exercised here since the other country scores need not be correct either. In a companion paper, McIntosh (2017b), it was shown that the PISA rankings of Canada and Finland in 2012 were not correct. PISA ranked them equally but using the methods outlined here gave Finland a large advantage over Canada in all seven Mathematics clusters.

Discussion and Conclusions

The first issue to be discussed is why PISA results are not confirmed by the methods used here. The methodology employed here differs from that used by PISA in three important ways. First, the normalization follows a Verhelst-Glas configuration rather than making the difficulty parameters sum to zero. Secondly, the two parameter logistic model is used instead of the Rasch model with $\alpha=1$; this is more general and allowing item specific dispersion parameters fits the data better and is supported by likelihood ratio tests. Third, the two years are pooled and the DIF minimizing normalization is used in the estimation procedure. Fourth, there are no imputations used in the production of Θ^*. In PISA mathematics scores are computed for all respondents whether they actually answered mathematics item or not. An imputation procedure used by PISA, due to Mislevy (1991), in which scores are random draws from the conditional Joint distribution of mathematics, reading, and science. It would appear that this imputation procedure is a major contributor to the PISA result. Using imputed mathematics scores which are based on all three subjects appears to obscure the true difference in mathematical performance across the two years. The fact that restricted PISA scores are higher in 2012 than 2009 suggests that the other two differences in procedure are not essential in getting the correct results although the year differences are less pronounced in the two parameter logistic model with the Verhelst-Glas normalization.

All clusters are estimated with models involving regressors. Means of the variables representing respondent characteristics are similar across the two years with the exception that there are fewer respondents in 2012 whose father has a university degree. There are some differences in the regression coefficients but they are not large and do not uniformly favour one year over the other. Score differences across the two years are not much influenced by the presence of regressors in the model.

PISA produces what it calls an ability or proficiency score. This is an inaccurate description. Because of the normalization used these scores also depend on the difficulty of the items. However, this is not the only problem. θ_v^* is a pure measure of ability; it is not affected by how difficult the items were to answer but it is not sufficient when the objective of the test is to measure the effectiveness of the educational system. Both average item difficulty and average ability are needed to make an appropriate assessment of the effects of educational policy on educational performance.

Tables

Table 1. Question Performance and Difficulty Parameter Estimates For Cluster 1: 2009 and 2012

	2009		2012	
Question	$\bar{y}_j$	δ_j	$\bar{y}_j$	δ_j
M033Q01	0.80	-1.43 (0.13)	0.78	-1.16 (0.12)
M034Q01T	0.47	1.0	0.48	1.0
M155Q01	0.76	-0.39 (0.06)	0.75	-0.42 (0.06)
M155Q02D	0.84	-0.82 (0.07)	0.80	-0.82 (0.08)
M155Q03D	0.39	1.37 (0.04)	0.37	1.49 (0.04)
M155Q04T	0.63	0.13 (0.05)	0.64	-0.01 (0.06)
M411Q01	0.59	0.51 (0.04)	0.59	054 (0.04)
M411Q02	0.51	0.85 (0.04)	0.52	0.82 (0.04)
M442Q02	0.45	1.10 (0.04)	0.44	1.24 (0.04)
M462Q01D	0.06	3.44 (0.17)	0.07	3.70 (0.20)
M474Q01	0.72	-0.39 (0.06)	0.75	-0.41 (0.06)
M803Q01T	0.36	0.88 (0.04)	0.37	0.85 (0.04)

Table note: $\bar{y}_j$ is the sample average of the item success rate.

Table 2. PISA, Raw Item, PISA P* Scores, and Average Difficulty Parameter Estimates: 2009-2012

Cluster	Sample Size	PISA PV Score	Raw Item Score	P* Score	Average Item Difficulty
2009	6971	517.3	6.45	496.4	0.52
1		(1.31)	(0.04)	(1.44)	(0.03)
2012	6258	535.5	6.70	521.7	0.57
		(1.25)	(0.03)	(0.80)	(0.03)
2009	6785	513.4	6.32	505.3	-0.31
2		(1.37)	(0.03)	(1.52)	(0.04)
2012	7012	536.4	6.64	512.0	-0.67
		(0.81)	(0.03)	(0.70)	(0.07)
2009	7003	521.1	5.20	506.1	0.88
3		(1.22)	(0.03)	(1.50)	(0.03)
2012	6124	525.9	5.29	512.1	0.87

(0.81)	(0.03)	(1.13)	(0.03)

Table 3. Amended Swamanathan-Rogers Test Statistics Based on Equation (14)

Cluster	Anchor Item	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$
1	M034Q02T	-0.059 (0.162)	0.074 (0.161)
2	M828Q03	0.249 (0.279)	-0.590 (0.614)
3	M564Q02	-0.683 (0.601)	0.792 (0.743)

References.

Adams, R.J. & Wu, Margaret L (2007). The Mixed-Coefficient Multinomial Logit Model: A Generalized Form of the Rasch Model. Ch. 4 in von Davier and Carstensen (2007).

Aitkin, Murray & Aitkin, Irit (2011). Statistical Modeling of the National Assessment of Educational Progress. Statistics for Social and Behavioral Sciences, 23. Springer.

Fischer, Gerhard H. & Molenaar, Ivo W. (1995). Rasch Models; Foundations, Recent Developments, and Applications, Springer Verlag, New York.

Goldstein, Harvey, Gerard Bonnet, and Thierry Rocher (2007). Multilevel Structural Equation Models for the Analysis of Comparative Data on Educational Performance. Journal of Educational and Behavioral Statistics, 32: 252-286.

Heckman, James J. & Singer, Burton (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. Econometrica, 52: 271-320.

Hopmann, S, G. Brinek and M. Retzl (2007). PISA According to PISA. Lit-Verlag, Vienna.

Kreiner, Svend & Christensen, Karl Bang (2013). Analysis of Model Fit and Robustness. A New Look at The PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. On line in Psychometrika.

Lindsay, B. G, Clogg, C.L. & Crego, J. (1991). Semiparametric Estimation in the Rasch Model and Related Exponential Response models, Including a Simple Latent Class Model For Item Analysis. Journal of the American Statistical Association, 86: 96-107.

Lord, Frederic M. & Novick, Melvin R. (1968). Statistical Theories of Mental Test Scores, Addison -Wesley, Reading Massachusetts USA.

McIntosh, J. (2017a). Normalization rules for item response models: Simulation results. Communications in Statistics: Simulation and Computation. https://doi.org/10.1080/03610918.2017.1371746

------------- *(2017b).* PISA Country Rankings Valid? Results For Canada and Finland. Scandinavian Journal of Educational Research. https://doi.org/10.1080/03313831.2017.1420687.

------------- (2018). Dealing With Differential Item Functioning In Item Response Models. Unpublished discussion paper, Concordia University.

Mislevy R. J. (1991) Randomized-Based Inference About Latent Variables From Complex Samples. Psychometrika 56: 177-196.

OECD (2010). PISA 2009 Results: What Students Know and Can Do, Student performance in Reading, Mathematics and Science. Vol. I. Paris.

-------- (2012), PISA 2009 Technical Report, PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en

-------- (2014) PISA 2012 Results: What Students Know and Can Do Student Performance in Mathematics, Reading and Science. Vol. 1. Paris

Prais, S. J. (2003). Cautions on OECD's Recent Educational Survey (PISA). Oxford Review of Education 29: 139-163.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests.

Verhelst, Norman D. & Glas, C. A. W. (1995). `The One Parameter Logistic Model' Ch. 12 in Fischer and Molenaar (1995).

Wuttke, Joachim (2007). Uncertainties and Biases in PISA. In Hopmann et al (2007, ch. 10).

--------------- (2011). PISA & Co A Critical Bibliography. Online version url: http//www.messen-und-deuten.de/pisa/biblio.htm.