The OECD's Programme for International Student Assessment (PISA) Study: A Review of Its Basic Psychometric Concepts

Ali Ünlü, Daniel Kasper, Matthias Trendtel, and Michael Schurig

Abstract The Programme for International Student Assessment (PISA; e.g., OECD, Sample tasks from the PISA 2000 assessment, 2002a; OECD, Learning for tomorrow's world: first results from PISA 2003, 2004; OECD, PISA 2006: Science competencies for tomorrow's world, 2007; OECD, PISA 2009 Technical Report, 2012) is an international large scale assessment study that aims to assess the skills and knowledge of 15-year-old students, and based on the results, to compare education systems across the participating (about 70) countries (with a minimum number of approx. 4,500 tested students per country). Initiator of this Programme is the Organisation for Economic Co-operation and Development (OECD; www.pisa.oecd.org). We review the main methodological techniques of the PISA study. Primarily, we focus on the psychometric procedure applied for scaling items and persons. PISA proficiency scale construction and proficiency levels derived based on discretization of the continua are discussed. For a balanced reflection of the PISA methodology, questions and suggestions on the reproduction of international item parameters, as well as on scoring, classifying and reporting, are raised. We hope that along these lines the PISA analyses can be better understood and evaluated, and if necessary, possibly be improved.

A. Ünlü (✉) · D. Kasper · M. Trendtel · M. Schurig
Chair for Methods in Empirical Educational Research, TUM School of Education, and Centre for International Student Assessment (ZIB), Technische Universität München, Arcisstrasse 21, 80333 Munich, Germany
e-mail: ali.uenlue@tum.de; daniel.kasper@tum.de; matthias.trendtel@tum.de; michael.schurig@tum.de

M. Spiliopoulou et al. (eds.), *Data Analysis, Machine Learning and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-01595-8_45,

1 Introduction

PISA is an international large scale educational assessment study conducted by member countries of the OECD (2001, 2002a, 2004, 2007, 2010) and investigates how well 15-year-old students approaching the end of compulsory schooling are prepared to meet the challenges of today's knowledge societies (OECD 2005, 2012). The study does not focus on the students' achievement regarding a specific school curriculum but rather aims at measuring the students' ability to use their knowledge and skills to meet real-life challenges (OECD 2009a).

PISA started in 2000 and takes place every 3 years. Proficiencies in the domains reading, mathematics, and science are assessed. In each assessment cycle, one of these domains is chosen as the major domain under fine-grained investigation; reading in 2000, followed by mathematics in 2003, science in 2006, etc. The definitions of the domains can be found in the respective *assessment frameworks* (e.g., OECD 2009a). In addition to these domains, further competencies may also be assessed by a participating OECD member; for example digital reading in 2009 (OECD 2012). Besides the actual test in PISA, student and school questionnaires are used to provide additional background information (e.g., about the socio-economic status of a student). In PISA 2009 for instance, in addition to these questionnaires, in 14 countries parents were asked to fill in an optional questionnaire. The background information are used as so-called conditioning variables for the scaling of the PISA cognitive (i.e., test) data.

The number of countries (and economies) participating in PISA continues to increase (e.g., 32 and 65 countries for PISA 2000 and 2009, respectively). In each participating country, a sample of at least 150 schools (or all schools) were drawn. In each participating school, 35 students were drawn (in schools with less than 35 eligible students, all students were selected).

The PISA study involves a number of technical challenges; for example, the development of test design and measurement instruments, of survey and questionnaire scales. Accurate sampling designs, including both school sampling and student sampling, must be developed. The multilingual and multicultural nature of the assessment must be taken into account, and various operational control and validation procedures have to be applied. Focused on in this paper, the scaling and analysis of the data require sophisticated psychometric methods, and PISA employs a scaling model based on *item response theory* (IRT; e.g., Adams et al. 1997; Fischer and Molenaar 1995; van der Linden and Hambleton 1997). The proficiency scales and levels, which are the basic tool in reporting PISA outcomes, are derived through IRT analyses.

The PISA *Technical Report* describes those methodologies (OECD 2002b, 2005, 2009b, 2012). The description is provided at a level that allows for review and, *potentially*, replication of the implemented procedures. In this paper, we recapitulate the scaling procedure that is used in PISA (Sect. 2). We discuss the construction of proficiency scale and proficiency levels and explain how the results are reported and interpreted in PISA (Sect. 3). We comment on whether information provided in the

Technical Report is sufficient to replicate the sampling and scaling procedures and central results for PISA, on classification procedures and alternatives thereof, and on other, for instance more automated, ways for reporting in the PISA Technical Report (Sect. 4). Overall, limitations of PISA and some reflections and suggestions for improvement are described and scattered throughout the paper.

2 Scaling Procedure

To scale the PISA cognitive data, the *mixed coefficients multinomial logit model* (MCMLM; Adams et al. 1997) is applied (OECD 2012, Chap. 9). This model is a generalized form of the Rasch model (Rasch 1980) in IRT. In the MCMLM, the items are characterized by a *fixed* set of unknown parameters, $\boldsymbol{\xi}$, and the student outcome levels, the latent random variable $\boldsymbol{\theta}$, are assumed to be *random effects*.

2.1 Notation

Assume I items (indexed $i = 1, \ldots, I$) with $K_i + 1$ possible response categories $(0, 1, \ldots, K_i)$ for an item i. The vector-valued random variable, for a sampled person, $\boldsymbol{X_i'} = \left(X_{i1}, X_{i2}, \ldots, X_{iK_i}\right)$ of order $1 \times K_i$, with $X_{ij} = 1$ if the response of the person to item i is in category j, or $X_{ij} = 0$ otherwise, indicates the $K_i + 1$ possible responses of the person to item i. The zero category of an item is denoted with a vector consisting of zeros, making the zero category a reference category, for model identification. Collecting the $\boldsymbol{X_i'}$'s together into a vector $\boldsymbol{X'} = \left(\boldsymbol{X_1'}, \boldsymbol{X_2'}, \ldots, \boldsymbol{X_I'}\right)$ of order $1 \times t$ ($t = K_1 + \cdots + K_I$) gives the response vector, or response pattern, of the person on the whole test.

In addition to the response vector $\boldsymbol{X}$ (person level), assume an $1 \times p$ vector $\boldsymbol{\xi'} = \left(\xi_1, \xi_2, \ldots, \xi_p\right)$ of p parameters ($p \geq I$) describing the I items. These are often interpreted as the items' difficulties. In the response probability model, linear combinations of these parameters are used, to describe the empirical characteristics of the response categories of each item. To define these linear combinations, a set of design vectors $\boldsymbol{a_{ij}}$ ($i = 1, \ldots, I$; $j = 1, \ldots, K_i$), each of length p, can be collected to form an $p \times t$ design matrix $\boldsymbol{A'} = \left(\boldsymbol{a_{11}}, \ldots, \boldsymbol{a_{1K_1}}, \boldsymbol{a_{21}}, \ldots, \boldsymbol{a_{2K_2}}, \ldots, \boldsymbol{a_{I1}}, \ldots, \boldsymbol{a_{IK_I}}\right)$, and the linear combinations are calculated by $\boldsymbol{A\xi}$ (of order $t \times 1$). In the multidimensional version of the model it is assumed that $D \geq 2$ latent traits underlie the persons' responses. The scores of the individuals on these latent traits are collected in the $D \times 1$ vector $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_D\right)'$, where the θ's are real-valued and often interpreted as the persons' abilities.

In the model also the notion of a response score b_{ijd} is introduced, which gives the performance level of an observed response in category j of item i with respect to dimension d ($d = 1, \ldots, D$). For dimension d and item i, the response scores

across the K_i categories of item i can be collected in an $K_i \times 1$ vector $\boldsymbol{b_{id}} = \left(b_{i1d}, \ldots, b_{iK_i d}\right)'$ and across the D dimensions in the $K_i \times D$ scoring sub-matrix $\boldsymbol{B_i} = \left(\boldsymbol{b_{i1}}, \ldots, \boldsymbol{b_{iD}}\right)$. For all items, the response scores can be collected in an $t \times D$ scoring matrix $\boldsymbol{B} = \left(\boldsymbol{B_1'}, \ldots, \boldsymbol{B_I'}\right)'$.

2.2 MCMLM

The probability $\Pr(X_{ij} = 1; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi}|\boldsymbol{\theta})$ of a response in category j of item i, given an ability vector $\boldsymbol{\theta}$, is $\exp(\boldsymbol{b_{ij}\theta} + \boldsymbol{a_{ij}'\xi})/(1 + \sum_{q=1}^{K_i}\exp(\boldsymbol{b_{iq}\theta} + \boldsymbol{a_{iq}'\xi})$, where $\boldsymbol{b_{iq}}$ is the qth row of the corresponding matrix $\boldsymbol{B_i}$, and $\boldsymbol{a_{iq}'}$ is the $(\sum_{l=1}^{i-1} K_l + q)$th row of the matrix $\boldsymbol{A}$. The *conditional item response model* (conditional on a person's ability $\boldsymbol{\theta}$) then can be expressed by $f_x(\boldsymbol{x}; \boldsymbol{\xi}|\boldsymbol{\theta}) = \exp\left[\boldsymbol{x'}\left(\boldsymbol{B\theta} + \boldsymbol{A\xi}\right)\right] / \sum_{\boldsymbol{z}} \exp\left[\boldsymbol{z'}\left(\boldsymbol{B\theta} + \boldsymbol{A\xi}\right)\right]$, where $\boldsymbol{x}$ is a realization of $\mathbf{X}$ and $\sum$ is over of all possible response vectors $\boldsymbol{z}$.

In the conditional item response model, $\boldsymbol{\theta}$ is given. The unconditional, or marginal, item response model requires the specification of a density, $f_\theta(\boldsymbol{\theta})$. In the PISA scaling procedure, students are assumed to have been sampled from a multivariate normal population with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, that is, $f_\theta(\boldsymbol{\theta}) = ((2\pi)^D\,|\boldsymbol{\Sigma}|)^{-1/2}\exp\left[-\left(\boldsymbol{\theta} - \boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta} - \boldsymbol{\mu}\right)/2\right]$. Moreover, this mean vector is parametrized, $\boldsymbol{\mu} = \boldsymbol{\Gamma' w}$, so that $\boldsymbol{\theta} = \boldsymbol{\Gamma' w} + \boldsymbol{e}$, where $\boldsymbol{w}$ is an $u \times 1$ vector of u fixed and known background values for a student, $\boldsymbol{\Gamma}$ is an $u \times D$ matrix of regression coefficients, and the error term $\boldsymbol{e}$ is $\boldsymbol{N(0, \Sigma)}$. In PISA, $\boldsymbol{\theta} = \boldsymbol{\Gamma' w} + \boldsymbol{e}$ is referred to as *latent regression*, and $\boldsymbol{w}$ comprises the so-called *conditioning variables* (e.g., gender, grade, or school size). This is the *population model*.

The conditional item response model and the population model are combined to obtain the *unconditional*, or *marginal*, *item response model*, which incorporates not only performance on the items but also information about the students' background: $f(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Gamma}, \boldsymbol{w}, \Sigma) = \int_\theta f_x(\boldsymbol{x}; \boldsymbol{\xi}|\boldsymbol{\theta})\, f_\theta(\boldsymbol{\theta}; \boldsymbol{\Gamma}, \boldsymbol{w}, \Sigma) d\boldsymbol{\theta}$. The parameters of this MCMLM are $\boldsymbol{\Gamma}$, Σ, and $\boldsymbol{\xi}$. They can be estimated using the software *ConQuest*® (Wu et al. 1997; see also Adams et al. 1997).

Parametrizing a multivariate mean of a prior distribution for the person ability can also be applied to the broader family of *multidimensional item response models* (e.g., Reckase 2009). Alternative models capable of capturing the multidimensional aspects of the data, and at the same time, allowing for the incorporation of covariate information are *explanatory item response models* (e.g., De Boeck and Wilson 2004). The scaling procedure in PISA may be performed using those models. In further research, it would be interesting to compare the different approaches to scaling the PISA cognitive data.

2.3 Student Score Generation

For each student (response pattern) it is possible to specify a posterior distribution for the latent variable $\boldsymbol{\theta}$, which is given by $h_\theta(\boldsymbol{\theta}; \boldsymbol{w}, \boldsymbol{\xi}, \boldsymbol{\Gamma}, \Sigma|\boldsymbol{x}) = f_x(\boldsymbol{x}; \boldsymbol{\xi}|\boldsymbol{\theta}) f_\theta(\boldsymbol{\theta}; \boldsymbol{\Gamma}, \boldsymbol{w}, \Sigma) / \int_\theta f_x(\boldsymbol{x}; \boldsymbol{\xi}|\boldsymbol{\theta}) f_\theta(\boldsymbol{\theta}; \boldsymbol{\Gamma}, \boldsymbol{w}, \Sigma)$. Estimates for $\boldsymbol{\theta}$ are random draws from this posterior distribution, and they are called *plausible values* (e.g., see Mislevy 1991).

Plausible values are drawn in PISA as follows. M vector-valued random deviates $(\boldsymbol{\varphi_{mn}})_{m=1,\ldots,M}$ are sampled from the parametrized multivariate normal distribution, for each individual n. For PISA, the value $M = 2{,}000$ has been specified (OECD 2012). These vectors are used to approximate the integral in the equation for the posterior distribution, using Monte-Carlo integration: $\int_\theta f_x(\boldsymbol{x}; \boldsymbol{\xi}|\boldsymbol{\theta}) f_\theta(\boldsymbol{\theta}; \boldsymbol{\Gamma}, \boldsymbol{w}, \Sigma) d\boldsymbol{\theta} \approx \frac{1}{M} \sum_{m=1}^{M} f_x(\boldsymbol{x_n}; \boldsymbol{\xi}|\boldsymbol{\varphi_{mn}}) = \Im$. The values $p_{mn} = f_x(\boldsymbol{x_n}; \boldsymbol{\xi}|\boldsymbol{\varphi_{mn}}) f_\theta(\boldsymbol{\varphi_{mn}}; \boldsymbol{\Gamma}, \boldsymbol{w}, \Sigma)$ are calculated, and the set of pairs $(\boldsymbol{\varphi_{mn}}, p_{mn}/\Im)_{m=1,\ldots,M}$ can be used as an approximation of the posterior density; and the probability that $\boldsymbol{\varphi_{jn}}$ could be drawn from this density is given by $q_{jn} = p_{jn} / \sum_{m=1}^{M} p_{mn}$. L uniformly distributed random numbers $(\eta_i)_{i=1}^{L}$ are generated; and for each random draw, the vector, $\boldsymbol{\varphi_{i_0 n}}$, for which the condition $\sum_{s=1}^{i_0-1} q_{sn} < \eta_i \leq \sum_{s=1}^{i_0} q_{sn}$ is satisfied, is selected as a *plausible vector*.

A computational question that remains unclear at this point concerns the mode of drawing plausible values. A perfect reproduction of the generated PISA plausible values is not possible. It also remains unclear which of the plausible values (for a student, generally five values are generated for each dimension), if the means of those values, or if even aggregations of individual results (computed one for each plausible value), were used for "classifying" individuals into the proficiency levels.

The MCMLM is fitted to each national data set, based on the international item parameters and national conditioning variables. However, the random sub-sample of students across the participating nations and economies used for estimating the parameters, is not identifiable (e.g., OECD 2009b, p. 197). Hence, the item parameters cannot be reproduced with certainty as well.

3 Proficiency Scale Construction and Proficiency Levels

In addition to plausible values, PISA also reports proficiency (scale) levels. The proficiency scales developed in PISA do not describe what students at a given level on the PISA "*performance* scale" *actually did* in a test situation, rather they describe what students at a given level on the PISA "*proficiency* scale" *typically know and can do*. Through the scaling procedure discussed in previous section, it is possible to locate student ability and item difficulty on "performance continua" $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, respectively. These continua are discretized in a specific way to yield the proficiency scales with their discrete levels.

Level	Score points on the PISA scale
6	Higher than 698.32
5	Higher than 625.61 and less than or equal to 698.32
4	Higher than 552.89 and less than or equal to 625.61
3	Higher than 480.18 and less than or equal to 552.89
2	Higher than 407.47 and less than or equal to 480.18
1a	Higher than 334.75 and less than or equal to 407.47
1b	262.04 to less than or equal to 334.75

Fig. 1 Print reading proficiency scale and levels (taken from OECD 2012, p. 266). PISA scales were linear transformations of the natural logit metrics that result from the PISA scaling procedure. Transformations were chosen so that mean and standard deviation of the PISA scores were 500 and 100, respectively (OECD 2012, p. 143)

The methodology to construct proficiency scales and to associate students with their levels was developed and used for PISA 2000, and it was essentially retained for PISA 2009. In the PISA 2000 cycle, defining the proficiency levels progressed in two broad phases. In the first phase, a substantive analysis of the PISA items in relation to the aspects of literacy that underpinned each test domain was carried out. This analysis produced a detailed summary of the cognitive demands of the PISA items, and together with information about the items' difficulty, descriptions of increasing proficiency. In the second phase, decisions about where to set *cut-off points* to construct the levels and how to associate students with each level were made.

For implementing these principles, a method has been developed that links three variables (for details, see OECD 2012, Chap. 15): the expected success of a student at a particular proficiency level on items that are uniformly spread across that level (proposed is a minimum of 50 % for students at the bottom of the level and higher for other students at that level); the width of a level in the scale (determined largely by substantive considerations of the cognitive demands of items at that level and observations of student performance on the items); and the probability that a student in the middle of the level would correctly answer an item of average difficulty for this level (referred to as the "RP-value" for the scale, where "RP" indicates "response probability").

As an example, for print reading in PISA 2009, seven levels of proficiency were defined; see Fig. 1.

A description of the sixth proficiency level can be found in Fig. 2.

The PISA study provides a basis for international collaboration in defining and implementing educational policies. The described proficiency scales and the distributions of proficiency levels in the different countries play a central role in the reporting of the PISA results. For example, in all international reports the percentage of students performing at each of the proficiency levels is presented (see OECD 2001, 2004, 2007, 2010). Therefore, it is essential to determine the proficiency scales and levels reliably.

Level	Lower score limit	Percentage of students able to perform tasks at this level or above	Characteristics of tasks
6	698	0.8% of students across the OECD can perform tasks at least at Level 6 on the reading scale	Tasks at this level typically require the reader to make multiple inferences, comparisons and contrasts that are both detailed and precise. They require demonstration of a full and detailed understanding of one or more texts and may involve integrating information from more than one text. Tasks may require the reader to deal with unfamiliar ideas, in the presence of prominent competing information, and to generate abstract categories for interpretations. Reflect and evaluate tasks may require the reader to hypothesise about or critically evaluate a complex text on an unfamiliar topic, taking into account multiple criteria or perspectives, and applying sophisticated understandings from beyond the text. A salient condition for access and retrieve tasks at this level is precision of analysis and fine attention to detail that is inconspicuous in the texts.

Fig. 2 Summary description of the sixth proficiency level on the print reading proficiency scale (taken from OECD 2012, p. 267)

Are there alternatives? It is important to note that specification of the proficiency levels and classification based on the proficiency scale depend on qualitative expert judgments. Statistical statements about the reliability of the PISA classifications (e.g., using numerical misclassification rates) are not possible in general, in the sense of a principled psychometric theory. Such a theory can be based on (order) restricted latent class models (see Sect. 4).

4 Conclusion

The basic psychometric concepts underlying the PISA surveys are elaborate. Complex statistical methods are applied to simultaneously scale persons and items in categorical large scale assessment data based on latent variables.

A number of questions remain unanswered when it comes to trying to replicate the PISA scaling results. For example, for student score generation international item parameters are used. These parameters are estimated based on a sub-sample of the international student sample. Although all international data sets are freely available (www.oecd.org/pisa/pisaproducts), it is not evident which students were contained in that sub-sample. It would have been easy to add a filter variable, or at least, to describe the randomization process more precisely. Regarding the reproduction of the plausible values it seems appropriate that, at least, the random number seeds are tabulated. It should also be reported clearly whether the plausible values themselves are aggregated before, for instance, the PISA scores are calculated, or whether the PISA scores are computed separately for any plausible value and aggregated. Indeed, the sequence of averaging may matter (e.g., von Davier et al. 2009).

An interesting alternative to the "two-step discretization approach" in PISA for the construction of proficiency scales and levels are psychometric model-based classification methods such as the *cognitive diagnosis models* (e.g., DiBello et al. 2007; Rupp et al. 2010; von Davier 2010). The latter are discrete latent variable

models (restricted latent class models), so no discretization (e.g., based on subjective expert judgments) is necessary, and classification based on these diagnostic models is purely statistical. We expect that such an approach may improve on the error of classification.

It may be useful to automatize the reporting in PISA. One way to implement that, is by utilizing *Sweave* (Leisch 2002). *Sweave* is a tool that allows to embed *R* code for complete data analyses in LaTeX documents. The purpose is to create *dynamic* reports, which can be updated *automatically* if data or analysis change. This tool can facilitate the reporting in PISA. Interestingly, different educational large scale assessment studies may then be compared, *heuristically*, data or text mining their Technical Reports.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Dibello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (pp. 979–1030). Amsterdam: Elsevier.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer.

Leisch, F. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. In W. Härdle & B. Rönz (Eds.), *Compstat 2002—Proceedings in Computational Statistics* (pp. 575–580). Heidelberg: Physica Verlag.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196.

OECD (2001). *Knowledge and skills for life. First results from the OECD Programme for International Student Assessment (PISA) 2000*. Paris: OECD Publishing.

OECD (2002a). *Sample tasks from the PISA 2000 assessment*. Paris: OECD Publishing.

OECD (2002b). *PISA 2000 Technical Report*. Paris: OECD Publishing.

OECD (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD Publishing.

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

OECD (2007). *PISA 2006: Science competencies for tomorrow's world*. Paris: OECD Publishing.

OECD (2009a). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science*. Paris: OECD Publishing.

OECD (2009b). *PISA 2006 Technical Report*. Paris: OECD Publishing.

OECD (2010). *PISA 2009 results: Overcoming social background—Equity learning opportunities and outcomes*. Paris: OECD Publishing.

OECD (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guilford Press.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8–28.

von Davier, M., Gonzales, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *Issues and Methodologies in Large-Scale Assessments, 2*, 9–37.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest®: Multi-aspect test software* [Computer program manual]. Camberwell: Australian Council for Educational Research.