Manfred Prenzel
Mareike Kobarg
Katrin Schöps
Silke Rönnebeck *Editors*

Research on PISA

Research Outcomes of the PISA Research Conference 2009

Springer

Research on PISA

Manfred Prenzel • Mareike Kobarg
Katrin Schöps • Silke Rönnebeck
Editors

Research on PISA

Research Outcomes of the PISA Research Conference 2009

Springer

Editors

Manfred Prenzel
School of Education
Technische Universität München
München, Germany

Mareike Kobarg
Leibniz Institute for Science
and Mathematics Education
IPN, Kiel, Germany

Katrin Schöps
Leibniz Institute for Science
and Mathematics Education
IPN, Kiel, Germany

Silke Rönnebeck
Leibniz Institute for Science
and Mathematics Education
IPN, Kiel, Germany

SPONSORED BY THE

Federal Ministry
of Education
and Research

ISBN 978-94-007-4457-8 ISBN 978-94-007-4458-5 (eBook)
DOI 10.1007/978-94-007-4458-5
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012938229

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Acknowledgements

This book summarizes central outcomes of the PISA Research Conference 2009 in Kiel, Germany.

We would like to thank the Norwegian Ministry of Education and Research who supported the preparation of the conference as well as the German Federal Ministry of Education and Research for their generous contribution to the conference itself.

Furthermore we would like to thank Barbara Drechsel, Katrin Gutzmann and Gráinne Newcombe for organizing the conference that lead to this book, our reviewers Claus H. Carstensen, Inger Marie Dalehefte, Barbara Drechsel, Timo Ehmke, Aiso Heinze, Nina Jude, Olaf Köller, Michael Leucht, Alexander Robitzsch, Christine Sälzer, Kerstin Schütte, Päivi Taskinen and Jörg Wittwer for their helpful comments on different chapters, Cornelia Gerigk for proofreading a number of chapters and Verena Hane and Emmanuel Kaiser for editing and redrawing the figures in this book.

Contents

Part III Context Related Research

Part IV Research on Trends in PISA

Contributors

Domenico Angelone Institute for Educational Evaluation, Associated Institute of the University of Zurich, Zürich, Switzerland

Eduardo Backhoff Institute for Educational Research and Development, University of Baja California, Ensenada, BC, Mexico

Werner Blum University of Kassel, Institute for Mathematics, Kassel, Germany

Claus H. Carstensen Institute of Psychology, University of Bamberg, Bamberg, Germany

Luis Angel Contreras-Niño Instituto de Investigación y Desarrollo Educativo, University of Baja California, Mexico, Ensenada, BC, Mexico

Matthias von Davier Research & Development Division, Educational Testing Service, Princeton, NJ, USA

John Dossey Department of Mathematics (Emeritus), Illinois State University, Tucson, USA

Harrie Eijkelhof Freudenthal Institute for Science and Mathematics Education, Utrecht University, Utrecht, The Netherlands

Andreas Frey Institute of Educational Science, Friedrich-Schiller-University Jena, Jena, Germany

Kylie Hillman Educational Monitoring and Research, Australian Council for Educational Research (ACER), Camberwell, Australia

Eckhard Klieme German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany

Mareike Kobarg Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany

Johanna Kordes Cito Institute for Educational Measurement, Arnhem, The Netherlands

Sheila Krawchuk Westat, Norcross, GA, USA

Ulf Kröhne Center for Research on Educational Quality and Evaluation, German Institute for International Educational Research, Frankfurt am Main, Germany

Christian Monseur Département Education et Formation, Université de Liège, Liège, Belgium

Urs Moser Institute for Educational Evaluation, Associated Institute of the University of Zurich, Zürich, Switzerland

Tarek Mostafa Centre for Learning and Life Chances in Knowledge Economies and Societies (LLAKES), Institute of Education, University of London, England, UK

Michael Neubrand Mathematics Education, Institute for Mathematics, Carl-von-Ossietzky-University, Oldenburg, Germany

Mogens Niss Mathematics and Mathematics Education, IMFUFA/NSM, Roskilde University, Roskilde, Denmark

Manfred Prenzel TU Muenchen School of Education, Susanne Klatten Endowed Chair for Empirical Educational Research, Schellingstr. 33, 80799 Munich, Germany

Silke Rönnebeck Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany

Keith Rust Westat, Rockville, MD, USA

Elwin Savelsbergh Freudenthal Institute for Science and Mathematics Education, Utrecht University, Utrecht, The Netherlands

Katrin Schöps Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany

Nicki-Nils Seitz Institute of Educational Science, Friedrich-Schiller-University Jena, Jena, Germany

Guillermo Solano-Flores School of Education, University of Colorado at Boulder, Boulder, CO, USA

Sue Thomson Educational Monitoring and Research, Australian Council for Educational Research (ACER), Camberwell, Australia

Ross Turner Australian Council for Educational Research (ACER), Camberwell, Australia

Daniel Urbach Psychometrics and Methodology, Australian Council for Educational Research (ACER), Camberwell, Australia

Kitty Williams Westat, Rockville, MD, USA

Trevor Williams Westat, Rockville, MD, USA

Mark Wilson Graduate School of Education, University of California, Berkeley, CA, USA

Introduction: Research on PISA, with PISA, and for PISA

Manfred Prenzel

The Purpose of PISA

The OECD "Programme for International Student Assessment" (PISA) is the most ambitious, comprehensive and best-known large-scale assessment in the field of education. Starting in 2000, every three years PISA measures knowledge and skills of students at the end of compulsory education. Assessing the competencies of 15 year old students in selected domains provides information on the quality of educational outcomes with respect to challenges of the so-called knowledge society. The participating countries expect findings from this international monitoring programme that may help to identify strengths and weaknesses of their educational system.

According to OECD (2006, p. 7) key features of PISA are:

- Its policy orientation, with design and reporting methods determined by the need of governments to draw policy lessons,
- Its innovative "literacy" concept, which is concerned with the capacity of students to apply knowledge and skills in key subject areas and to analyse, reason and communicate effectively as they pose, solve and interpret problems in a variety of situations,
- Its relevance to lifelong learning, which does not limit PISA to assessing students' curricular and cross-curricular competencies but also asks them to report on their own motivation to learn, their beliefs about themselves and their learning strategies,
- Its regularity, which will enable countries to monitor their progress in meeting key learning objectives,

M. Prenzel (✉)
TU Muenchen School of Education, Susanne Klatten Endowed Chair for Empirical Educational Research, Schellingstr. 33, 80799 Munich, Germany
e-mail: manfred.prenzel@tum.de

– Its contextualisation within the system of OECD education indicators, which examine the quality of learning outcomes, the policy levers and contextual factors that shape these outcomes, and the broader private and social returns to investments in education (OECD, 2006, p. 7).

The relevance of the findings from such a programme depends on the quality of the theoretical framework, the methods and statistical analysis. According to this, frameworks, methods and analysis have to be based on the state of the art in the appropriate fields of research. Otherwise the scientific community would immediately question the validity and credibility of the study, and accordingly the findings would become worthless for policy.

As a long-term programme PISA also must keep pace with the scientific progress. Compared to research in typical scientific contexts, however, PISA cannot implement ideas of "high-risk" research. The policy orientation compels a solid, and to some extent, conservative research strategy. Of course, cutting-edge approaches to assessment or to data analysis have to be considered, but PISA first of all needs proven methods that are broadly accepted and allow maintaining the comparability across countries and cycles.

Such considerations make it clear that in each new cycle PISA will provide reliable and representative data through its international assessments. Without doubt these findings will contribute to the wealth of scientific knowledge and lead to new insights – and new research questions as well. Basic research is not PISA's main purpose: It is also driven and controlled by political needs, and not by (pure) scientific interest. Nonetheless, PISA contributes to research, and depends on research. Some of the relationships between PISA and research will be analysed and discussed in the following.

A first look from a research perspective classifies PISA as a survey study with a cross-sectional and trend design, using e.g., representative random samples, standardized assessments, IRT scaling methods, transparent data analysis, etc. (Seidel & Prenzel, 2008). Such descriptions and classifications underline the quality of PISA in terms of typical research criteria. Experts in the field appreciate the methodological quality of PISA and take serious consideration of PISA data.

At the same time such classifications of the type of research also indicate serious limitations of the PISA approach, for example with respect to causal or prescriptive interpretations of findings. Due to the design, PISA itself does not provide sufficient evidence for causal explanations and also certainly not for scientifically approved measures or interventions in order to improve national educational systems. However, PISA provides excellent evidence to identify strengths and weaknesses of educational systems as well as examples of good practice in the sense of benchmarking. From a long-term perspective PISA has great significance as the programme allows tracking effects of measures that meanwhile have been implemented.

Limitations due to the design of studies are natural for researchers. In the context of PISA such limitations have to be communicated to policy and public. As PISA may identify severe problem areas in educational systems, researchers may feel motivated to do more and specific research that helps to go beyond such boundaries of the usual PISA design. Of course, additional research into such issues needs additional financing. In political contexts it must sometimes be stressed that PISA does not replace specific research programmes aiming to achieve theoretical insights in educational systems providing explanations and evidence for interventions and innovations. More often PISA findings will lead to serious questions that can only be answered on the basis of additional research programmes. Researchers will be especially interested if the findings lead to questions or issues that are as well relevant and challenging when taking into consideration recent theories in the field. A strong orientation of PISA to recent topics of research increases the likelihood that scientists will pick up issues and advance their research in such directions.

All in all, the quality of PISA depends on the current state of the art in a number of research fields (from domain specific assessment of competencies to teaching and learning processes, influences of context factors to issues of scaling and statistical modelling). PISA – as a programme already monitoring and comparing educational outcomes and systems on a high methodological standard – has the potential to become complemented and extended towards excellent research programmes. Additional research programmes can help improve PISA and to get more relevant information out of PISA – especially information that can help improve schools and educational systems.

Structure of Research on PISA

Since the first reports of PISA 2000 (e.g. OECD, 2001) thousands of scientific articles have been published on PISA (in November 2011 Google scholar listed approximately 160,000 articles). These articles differ, of course, in scientific quality. Only the smaller proportion of these papers is based on original analysis of PISA data going beyond the OECD reports. Especially these articles complement and greatly expand the OECD publications and make essential contributions to the scientific discourse on PISA.

Significant relevance can be attributed to the numerous national PISA reports, and especially to those that were are based on profound own analysis and – in several cases – on additional questionnaire or assessment data as well as on extensions of the national sample (for an overview of such reports, e.g.: https://mypisa.acer.edu.au/index.php?option=com_content&task=view&id=70&Itemid=446). Extended national reports (e.g. Bussiere, Knighton, & Pennock, 2007; Caygill & Sok, 2008; De Bortoli & Thomson, 2009; Hautamäki et al., 2008; Ho, 2008; Kjærnsli & Lie, 2004; Lie, Linnakyl, & Roe, 2003; Matti, 2009; Prenzel & Baumert, 2008; Schreiner & Schwantner, 2009) give an idea how different approaches can be used to combine ongoing PISA cycles with additional research questions.

In many cases extensions of the 'normal' PISA programme (e.g. by introducing additional items, scales, questionnaires or by oversampling with respect to regions, grade or student background) can be realized under certain conditions as so called 'national options'. Additional research linked to a PISA cycle as well as differentiated secondary analyses often are carried out in collaboration with universities or research institutes. Even more complex and demanding is the combination of PISA with follow-up or even longitudinal studies. A PISA cycle may also be used to explore innovative methods for assessment (e.g. technology-based assessments) or psychometric or statistical analysis.

Systematically, research extending or deepening the regular PISA cycles could be classified formally by design, instruments and statistical methods or by level of aggregation (from individual to group and system level), and concerning the contents by domains and constructs. Another important aspect (not only for the purpose of classification) refers to the funding of extensions and additional studies: Did the government fund the research or was it financed by national or international research councils/foundations? Accordingly it can be asked whether scientific committees have thoroughly reviewed the proposals and the reports.

The aims of research approaches in the periphery of PISA may also differ significantly from providing some additional descriptive information to extending the design for allowing causal analysis and interpretations. Additional research can be driven by the intention to critically question PISA and its findings – or to constructively help improve the design and the methods of PISA.

In the following a model for classifying research linked to PISA will be based on these considerations: Using the time structure of PISA cycles, three basic approaches can be differentiated: Research *for* PISA, research *with* PISA, and research *on* PISA.

- *Research for PISA* precedes a certain PISA cycle exploring and preparing (innovative) components of a PISA cycle in the future (e.g., assessment of social competencies; hands-on science assessment);
- *Research with PISA* proceeds in the course of a PISA cycle adding components (e.g., instruments, samples) in order to extend the scope, the significance and validity of PISA findings (e.g., oversampling of complete classrooms or other target groups; systematically combining PISA with national assessments);
- *Research on PISA* follows a PISA cycle doing additional, in depth-research of data from that cycle (e.g., secondary analysis of specific item groups; applying and testing new statistical approaches, exploring new approaches for trend analysis).

In order to specify what is meant with such approaches, an example for research with PISA will be presented in the following.

Research with PISA: How to Extend a PISA Cycle?

The consortium that was responsible for PISA 2003 in Germany applied for permission and for funding to significantly extend the national design of PISA (Prenzel et al., 2006). The Standing Conference of the Ministers of Education and Cultural Affairs of the Laender in the Federal Republic of Germany decided to support the project. Additional funding was raised from the German Research Foundation

(Deutsche Forschungsgemeinschaft) for a linked project on teacher competencies (so-called COACTIV-Study; cf. Kunter et al., 2007).

The study was labelled 'PISA I plus', because the international school sample served as a base to which components were added ('plus'):

- A random selection of two complete classes (9th grade) from all participating PISA schools,
- A second day of assessment where national tests and questionnaires were administered,
- An additional parent questionnaire and a teacher questionnaire (addressing a random sample of mathematics teachers from the school) were applied;
- The mathematics teachers of the selected classrooms were (as a part of COACTIV) extensively examined using different kinds of assessments and interviews;
- The most important addition, however, was a follow-up assessment of all selected classrooms and students 1 year later (grade 10) in mathematics and science.

In mathematics the assessment in 10th grade was based on the normal PISA assessment, but enriched with items on the curricular level of the higher grade. This approach allowed locating the performance of the students in grade 9 and 10 on the same scale (using a latent growth model).

A scatter-plot (Fig. 1) shows how much the students did learn in the course of 1 year in mathematics on that scale (Ehmke, Blum, Neubrand, Jordan, & Ulfig, 2006). The points represent about 6,000 students and show their performance in grade 9 and grade 10 on the PISA scale.

The diagonal line in the scatter-plot represents *no change* (no gain or loss) in mathematics performance between grade 9 and grade 10. The area above the diagonal line shows students who improved on the scale. On average, the performance gain was 25 points on the PISA scale. Obviously, a considerable proportion of students (34%) *did not improve* in mathematics over the course of 1 year (and having 4 h mathematics lesson per week on average). A decrease in performance was noted for 8% of the students. Analyses using different samples of items (e.g. literacy oriented vs. curriculum-oriented) led to similar results. The analysis did not provide evidence of differential growth, for examples depending on the previous mathematics performance or the type of school visited in the German tracking system. Only a small gender effect favouring girls was found (Ehmke et al., 2006).

For an interpretation of these finding it has to be considered that nearly all the students successfully passed the 9th as well as the10th grade, although it seemed that many did not improve their mathematics performance, in terms of the PISA tests (or a curriculum-oriented assessment).

Obviously, most of the students had successfully passed the frequent teacher made mathematics tests during the course of the school year and thus had received sufficient marks. So from the perspective of the students as well as of their teachers and parents, the year of schooling was quite successful. Using the information from the regular teacher-made tests they could more or less be satisfied or happy with the results indicating successful learning of mathematics. There was also no indication in the questionnaire data showing that students and teachers did not engage in learning

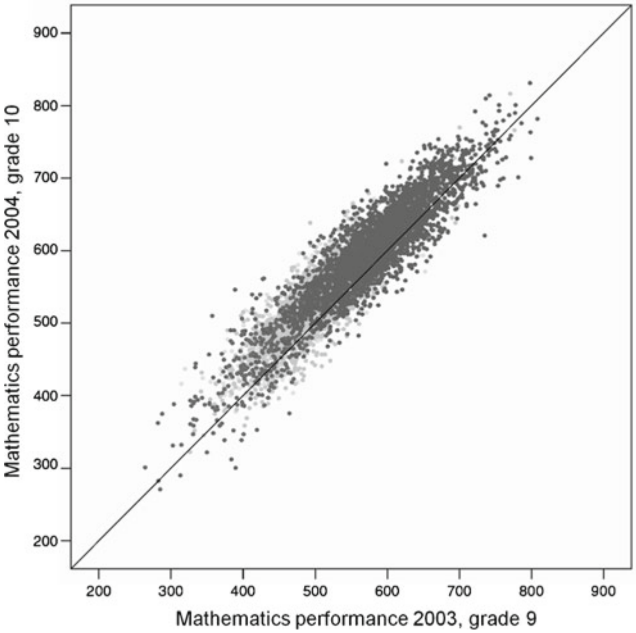


Fig. 1 Mathematics performance at the end of grade 9 and grade 10 (individual level, latent growth), N = 6020 (Ehmke et al., 2006, p. 74)

and teaching. From the perspective of a solid (delayed summative) assessment, however, a large proportion of students did not really improve. The findings provide evidence that the mathematics teaching and learning in German classrooms tends not to be very sustainable. The students all in all performed quite successful in the frequent teacher-made tests assessing material from a few weeks prior. However, the assessment some months after the instruction provides a very different picture. Many of the students showed only successful short-term learning. Findings on the sustainability of mathematics teaching help to interpret the performance of students from Germany in the international comparison.

The sampling of complete classes also allowed analysing the progress in mathematics performance on the aggregate level of classrooms and schools. The scatter plot in Fig. 2 shows the change in the mathematics performance aggregated on the level of (N = 275) classrooms (Ehmke et al., 2006).

On average relevant gains in the average classroom mathematics performance were found in 89% of the classes. More interesting are the differences between classes in the performance gains because the variance here offered manifold opportunities for more differentiated analysis of factors that, from a theoretical point of view, could make a difference. The analyses (cf. Prenzel et al., 2006) showed for example, that factors like teaching approach (quality of tasks, classroom management), student leisure activities and parental support have significant effects on mathematics performance in grade 10 (controlling the performance in grade 9).

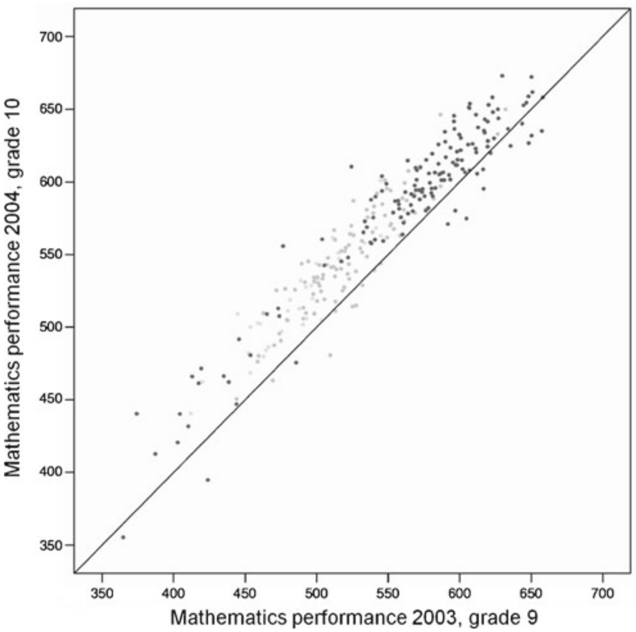


Fig. 2 Mathematics performance at the end of grade 9 and grade 10 (classroom level, latent growth), N = 275 (Ehmke et al., 2006, p. 77)

The COACTIV project linked to the follow-up design provided evidence on effects of relevant components of mathematics teachers' expertise on the performance of the students (e.g. Baumert et al., 2010; Kunter et al., 2007, 2008). Also analysis on the school level for example found effects of teacher composition and collaboration on mathematics and science performance.

The here presented follow-up study on PISA in Germany may serve as an example how the international assessment programme can be extended on a national level in the sense of 'research with PISA'. The report on PISA 2000 in Germany (Baumert et al., 2001) had triggered an intense political and public debate and raised many questions that could also be addressed in terms of research questions. Solid answers to a number of these questions could be expected from research (Prenzel, 2009). So it was consequent to extend the PISA 2003 design towards a follow-up study offering far more possibilities for the identification and interpretation of factors.

The international sample and assessments provide an excellent basis for enriching in-depth studies that help to explain and understand why students in a country perform at a certain level and at the same time contribute to the scientific knowledge in general. Extensions of the PISA design provide a win-win-situation for policy and research. So it makes sense for policy to invest additional money in research on PISA. On the other hand, researchers can demonstrate their interest in research linked to PISA if they apply for research funds allowing related studies. In Germany a group of researchers successfully applied for a Priority Programme from the

German Research Foundation dealing with issues of the educational quality of schools (Prenzel, 2007) shortly after the first PISA report. A number of projects in this programme were linked to PISA, such as the already mentioned COACTIV study on effects of teachers' competences on the mathematics performance of their students (Kunter et al., 2007). Another project started a longitudinal study on the development of mathematics competencies starting at the end of primary school and lasting up to age 15 when the regular PISA assessment takes place (Pekrun et al., 2007). In another study, a first attempt was made to assess the mathematics performance of parents with PISA tests and to relate the competencies of the parents to the learning progress of their children in school (Ehmke & Siegle, 2007). There were also projects using video techniques to complement the PISA questionnaire approach to teaching and learning, e.g. in science (Seidel et al., 2007). The findings of this study provided valuable insights for the construction of science teaching and learning scales in the international student questionnaire for PISA 2006 (cf. Kobarg et al., 2011).

Research on PISA: Some Expectations

The mentioned examples of studies illustrate different kinds of approaches to extend and complement PISA, an international assessment programme that most likely will be continued for decades. Because of its quality and continuity, PISA will presumably also serve educational policy in the future as a most relevant international monitoring programme. Also research in education can expect data on educational outcomes from PISA on a regular basis. It is in the interest of research in education to contribute to the quality and meaningfulness of the PISA design, instruments and data. Possibilities to combine large-scale assessments like PISA with more comprehensive research approaches should be strategically used in order to improve the scope and quality of data pertaining to educational systems as well as the theoretical understanding of relevant factors for successful education in school. Educational research that identifies problems in educational systems will get public visibility and attention by policy makers. Highest recognition will be given when educational research helps finding ways to solve such problems.

The PISA Research Conference in Kiel (Germany) was the first international meeting where scholars had the opportunity to exchange findings from their research on PISA and to discuss theoretical and methodological approaches as well as research strategies. Participants from research and from educational policy found a forum to exchange their views and to discuss research questions, priorities and possibilities for supporting research on PISA. The conference gave the impression of an emerging field of research on education because researchers in so many places around the world seem to be working enthusiastically on PISA related research. In line with that there was the general agreement that the first PISA Research Conference should be the starting point for a series of international meetings, not only for the exchange of findings and ideas, but also for the joint planning of bi- and multi-lateral international research co-operations.

References

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, J., & Weiß, M. (Eds.). (2001). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske & Budrich.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180.

Bussiere, P., Knighton, T., & Pennock, D. (2007). *Measuring up: Canadian results of the OECD PISA study: The performance of Canada's youth in science, reading and mathematics. 2006 first results for Canadians aged 15.* Ottawa, ON: Statistics Canada.

Caygill, R., & Sok, S. (2008). *PISA 2006. School context of Science achievement.* Wellington, New Zealand: Ministry of Education.

De Bortoli, L., & Thomson, S. (2009). *The achievement of Australia's Indigenous Students in PISA 2000–2006*. Camberwell, VA: ACER Press.

Ehmke, T., Blum, W., Neubrand, M., Jordan, A., & Ulfig, F. (2006). Wie verändert sich die mathematische Kompetenz von der neunten zur zehnten Klassenstufe? In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, J. Rost, & U. Schiefele (Eds.), *PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (S. 63–86). Muenster/New York: Waxmann.

Ehmke, T., & Siegle, T. (2007). How well do parents do in PISA? Results concerning the mathematical competency of parents. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (S. 61–77). Muenster/New York: Waxmann.

Hautamäki, J., Harjunen, E., Hautamäki, A., Karjalainen, T., Kupiainen, S., Laaksonen, S., et al. (2008). *PISA 2006 Finland. Analyses. Reflections. Explanations*. Helsinki, Finland: Finnish Ministry of Education.

Ho, E. S.-C. (2008). *The third Hongkong PISA report: PISA 2006. From PISA 2000 to PISA 2006.* Hongkong: HKPISA.

Kjærnsli, M., & Lie, S. (2004). PISA and scientific literacy: Similarities and differences between the Nordic countries. *Scandinavian Journal of Educational Research, 48*(3), 271–286.

Kobarg, M., Prenzel, M., Seidel, T., Walker, M., McCrae, B., Cresswell, J., & Wittwer, J. (2011). *An international comparison of science teaching and learning – Further results from PISA 2006*. Muenster, Germany/New York: Waxmann.

Kunter, M., Klusmann, U., Dubberke, T., Baumert, J., Blum, W., Brunner, M., et al. (2007). Linking aspects of teacher competence to their instruction. Results from the COACTIV Project. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (S. 32–52). Muenster/New York: Waxmann.

Kunter, M., Tsai, Y.-M., Klusmann, U., Brunner, M., Krauss, S., & Baumert, J. (2008). Students' and mathematics teachers' perceptions of teacher enthusiasm and instruction. *Learning and Instruction, 18*(5), 468–482.

Lie, S., Linnakylä, P., & Roe, A. (2003). *Northern Lights on PISA*. Oslo, Norway: Department of Teacher Education and School development. University of Oslo.

Matti, T. (Ed.). (2009). *Northern lights on PISA 2006. Differences and similarities in the Nordic countries*. Copenhagen, Denmark: Nordic Council of Ministers.

OECD (2001). *Knowledge and skills for life*. First results from PISA 2000. Paris: OECD

OECD (2006). *PISA – The OECD Programme for international Student Assessment ("PISA Brochure").* Paris: OECD. (http://www.oecd.org/dataoecd/51/27/37474503.pdf)

Pekrun, R., vom Hofe, R., Blum, W., Frenzel, A. C., Goetz, Th., & Wartha, S. (2007). Development of mathematical competencies in adolescence. The PALMA longitudinal study. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (S. 17–37). Muenster/New York: Waxmann.

Prenzel, M. (Ed.). (2007). *The educational quality of schools. Final report on the DFG priority programme.* Muenster, Germany/ New York: Waxmann.

Prenzel, M. (2009). Challenges facing the educational system. In European Science Foundation (Ed.), *Vital questions. The contributions of European Social Sciences* (pp. 30–34). Strasbourg, France: European Science Foundation.

Prenzel, M., & Baumert, J. (Eds.). (2008). Vertiefende Analysen zu PISA 2006. *Zeitschrift für Erziehungswissenschaft, Sonderheft 10*. Wiesbaden: VS Verlag.

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rost, J., & Schiefele, U. (Eds.). (2006). *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres*. Muenster, Germany/New York: Waxmann.

Schreiner, C., & Schwantner, U. (Eds.). (2009). *PISA 2006: Oesterreichischer Expertenbericht zum Naturwissenschafts-Schwerpunkt*. Graz, Austria: Leykam.

Seidel, T., & Prenzel, M. (2008). Large scale assessment. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts. State of the art and future prospects* (pp. 279–304). Goettingen, Germany: Hogrefe & Huber.

Seidel, T., Prenzel, M., Rimmele, R., Herweg, C., Kobarg, M., Schwindt, K., & Dalehefte, I. M. (2007). Science teaching and learning in German physics classrooms – Findings from the IPN video study. In M. Prenzel (Ed.), *The educational quality of schools. Final report on the DFG priority programme* (pp. 79–99). Muenster, Germany/New York: Waxmann.

Part I
Content Related Research

Introduction: Content Related Research on PISA

Werner Blum

In the PISA assessment, the domains are organized according to three interrelated aspects:

- The domain-specific contents that are targeted in the assessment problems of the respective domain,
- The domain-specific processes that describe what individuals ought to do and which capabilities they ought to activate in order to solve the problems,
- The contexts in which the problems are located.

Content related research as represented in this first part of the book may be geared toward any of these aspects. The key feature of content related research is a special focus on the *domain* whereas other relevant aspects such as the methodology used or relations to background variables are only secondary here.

There is a lot of empirical evidence that emphasising domain specific aspects is crucial both for the quality of teaching and for the effectiveness of research into teaching and learning. In the TIMSS video study, for instance, it turned out that mathematics lessons that were comparable concerning the topic, the lesson structure and the methods used, offered very different cognitive learning opportunities for students. This was revealed by in-depth analyses of the implemented mathematical tasks and the way these tasks were treated in the classroom (see, e.g., Klieme, Schümer, & Knoll, 2001; Stigler & Hiebert, 2004). The same results were reported by Kunter et al. (2006) in the context of the German supplement to the PISA study 2003/2004. Overall, there are meta-analyses showing that domain-specific aspects of instructional quality referring to content-related in-depth structures of lessons are substantially better predictors of students' learning progress than more general surface structures of lessons (see Baumert et al., 2010; Hattie, 2009; Seidel & Shavelson, 2007).

W. Blum
University of Kassel, Institute for Mathematics, D-34109, Kassel, Germany
e-mail: blum@mathematik.uni-kassel.de

A particularly important aspect is represented by the *tasks* that are covered in lessons or used for examinations. Mathematics is certainly the subject where tasks play the most dominant role (Christiansen & Walther, 1986), and also in the natural sciences the content is mostly represented by suitable tasks. By "task" we mean requiring students to work on a limited topic in a goal-oriented manner. For students, working on tasks is by far the most important activity in those subjects. This means tasks are the substance for the cognitive activities of learners. For teachers, tasks are a crucial element to orchestrate lessons and to clarify the aims of instruction as solving these tasks requires the competencies (see Niss, 2003) that students are to acquire. Research related to tasks is so important because of questions such as: How do teachers handle tasks in the classroom, how do learners deal with them, what inferences can be drawn from student solutions for diagnosing their competencies, how can tasks be constructed so as to foster certain (sub-)competencies, how can tasks be used for measuring certain facets of students' competencies, what aspects contribute to the difficulty of tasks, for instance: what roles do the context, the format, the wording play?

There is still a lack of empirical research into these questions, both for mathematics and for the sciences. PISA is a context where tasks are used to measure students' competence in various domains. This context can also be used for tackling some of the above mentioned research questions, for example identifying significant factors that influence the empirical difficulty of items. Moreover, the special focus of PISA leads to new research questions, for example concerning the suitability of different kinds of tasks for measuring purposes. As this PISA Research Conference has shown, there are several groups around the world that do research concerning tasks or, more generally, concerning the content component. A few of them are represented in this first part of the present volume.

The three papers in part I focus on different aspects. In the first paper by Eijkelhof, Kordes and Savelsbergh, with reference to the science tasks, Dutch results in PISA science are analyzed, and some conclusions are drawn for the Dutch science curriculum. The second paper by Turner, Dossey, Blum and Niss deals with the process component of PISA mathematics, that is, with the role of mathematical competencies for predicting item difficulty. The heart of this study is the analysis of the cognitive demands inherent in the PISA mathematics tasks. In the third paper, Neubrand reports on the conceptual framework for mathematics in the German supplement to PISA. Again, the tasks play a central role here. The three papers are described in more detail in the following.

Eijkelhof, Kordes and Savelsbergh start their contribution with a comparison of Dutch science education and the PISA science framework, and they find remarkable similarities with respect to science embedded in contexts. They then report on a detailed analysis carried out on the item level concerning relative strengths and weaknesses of Dutch students when solving various PISA science items, with a special emphasis on a distinction between students from general and from vocational secondary schools. These findings can mostly be explained by the relative emphasis of the item contexts and formats in Dutch science classrooms. In summary, the authors draw some conclusions for Dutch science education, in particular concerning the questions of how to improve vocational students' abilities to solve open-ended tasks and how to foster a more positive attitude of secondary students towards science.

Turner, Dossey, Blum and Niss report on investigations carried out by the international PISA Mathematics Expert Group over many years. In order to find an answer to the question of which factors influence item difficulty in PISA, they draw on a set of mathematical competencies that were originally developed in the Danish KOM project and from the beginning have been the conceptual core of the concept of mathematical literacy in PISA. The authors give operational descriptions of the six competencies and distinguish between four levels of cognitive demand in each competency. They then report on an extended empirical study in which all PISA items have been classified according to these levels for all competencies. It turned out that the psychometric quality of the coding was satisfactory and that about 70% of the variability in the PISA item difficulty data could be predicted by those competency related variables.

Neubrand starts with the observation that countries show considerably different results on certain PISA mathematics items which, on the country level, calls for a closer look into the PISA results from a mathematics education point of view. He then describes the model used as a conceptual framework for item construction and item analysis in the German supplement to PISA-2000. In particular, this model distinguishes between three "types of mathematical activities". The author shows that these types can also be used as a means to detect characteristic profiles of the mathematical achievement within Germany, that is, in the German federal states. The paper closes with a plea for differentiated assessment by categories adhering closer to the subject – such as the activity types mentioned above.

All three papers show that there are interesting domain specific and country specific questions in the PISA context that need particular attention. Above all, the three papers demonstrate the particular relevance of the content component in PISA, especially on the task level, and of further specific research into this area.

References

Baumert, J., Kunter, M., Blum, W., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American educational research journal, 47*(1), 133–180.

Christiansen, B., & Walther, G. (1986). Task and activity. In B. Christiansen, A. G. Howson, & M. Otte (Eds.), *Perspectives on mathematics education* (pp. 243–307). Dordrecht, the Netherlands: Reidel.

Hattie, J. A. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: "Aufgabenkultur" und Unterrichtsgestaltung im internationalen Vergleich. In E. Klieme & J. Baumert (Eds.), *TIMSS – Impulse für Schule und Unterricht: Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (pp. 43–57). Bonn, Germany: Bundesministerium für Bildung und Forschung.

Kunter, M., Dubberke, T., Baumert, J., Blum, W., Brunner, M., Jordan, A., et al. (2006). Mathematikunterricht in den PISA-Klassen 2004: Rahmenbedingungen, Formen und Lehr-Lernprozesse. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, et al. (Eds.), *PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (pp. 161–194). Münster, Germany: Waxmann.

Niss, M. (2003). Mathematical competencies and the learning of mathematics: The Danish KOM Project. In A. Gagatsis, & S. Papastavridis (Eds.), *3rd Mediterranean conference on mathematical education* (pp. 115–124). Athens: The Hellenic Mathematical Society.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499.

Stigler, J., & Hiebert, J. (2004). Improving mathematics teaching. *Educational Leadership, 61*(5), 12–17.

Chapter 1
Implications of PISA Outcomes for Science Curriculum Reform in the Netherlands

Harrie M.C. Eijkelhof, Johanna H. Kordes, and Elwin R. Savelsbergh

Abstract As the PISA 2006 results came out, the Netherlands briefly celebrated their 9th position in the overall country ranking for science. After that, interest in the PISA results rapidly declined. Nevertheless, there is sufficient reason to take a closer look at the PISA results, for instance because (a) our neighbours are catching up, (b) currently ambitious curriculum innovation programmes are being conducted in most of secondary education, and (c) fierce debates are going on about the merits of the proposed innovations. The pressing question is: are we heading in the right direction? To answer this question we additionally analysed PISA 2006 data, we identified strengths and weaknesses at the item level, and we analysed the student data for those specific items. As a reference for comparative analyses across countries, we used a relevant peer group of seven neighbouring countries. Main findings include that Dutch students do well on highly contextualized items, interpretation of graphs and *Knowledge of Science*. Dutch students perform relatively weak on items with low context and on multiple response items. In addition, Dutch students in secondary vocational education have specific difficulty in answering open-constructed response items. A major issue in the Netherlands is the low science attitudes and self-concept of students in secondary education. In view of those results, recent efforts to promote and improve science education might be well on track. However, we also identify some policy threats, especially when it comes to *Knowledge about Science*.

H.M.C. Eijkelhof, Ph.D. (✉) • E.R. Savelsbergh, Ph.D.
Freudenthal Institute for Science and Mathematics Education, Utrecht University,
P.O. Box 85.170, 3508 AD Utrecht, The Netherlands
e-mail: h.m.c.eijkelhof@uu.nl; e.r.savelsbergh@uu.nl

J.H. Kordes, Ph.D.
Cito Institute for Educational Measurement, P.O. Box 1034, 6801MG
Arnhem, The Netherlands
e-mail: johanna.kordes@cito.nl

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_1,

Keywords Attitudes toward science • Contexts in science education • Curriculum reform • International comparison

1.1 Introduction

Over the last 20 years in the Netherlands many curriculum development initiatives were taken to improve science education, both in lower and upper secondary education. Main aims of these efforts were to improve education by making topics more relevant to students, raising interest in science and technology and preparing students better for further education. Currently, new nation-wide curriculum reforms are going on, and the question arises whether the previous reforms were heading in the right direction, and what the next steps are to be taken. In this chapter, we draw on the outcomes of the third PISA-cycle (2006) to answer these questions.

We first compare the PISA 2006 Framework for Scientific Literacy (OECD, 2006) with past and current developments in science teaching in the Netherlands in order to formulate our expectations of Dutch students' strengths and weaknesses. Next, we explore what kinds of science items are relatively more easy or difficult for Dutch students in comparison to their peers in seven neighbouring European countries. We also focus on differences within the Netherlands in results between students in general and vocational education. After that, we compare the attitudes of Dutch students towards science with those of students in neighbouring countries. Finally we present some recommendations for Dutch science education based on our analyses.

1.2 Dutch Science Education and the PISA 2006 Scientific Literacy Framework

1.2.1 The PISA 2009 Scientific Literacy Framework

In preparation for the 2006 survey a science expert group led by Rodger Bybee prepared a framework in which the concept scientific literacy was defined (Bybee, McCrae, & Laurie, 2009). The framework used the following components: scientific contexts, scientific competencies, scientific knowledge and attitudes towards science.

The scientific contexts were framed within a variety of situations involving science and technology: health, natural resources, environment, hazards and frontiers of science and technology. They are each classified in situations at personal, social and global levels.

The scientific competencies are: *Identifying scientific issues*, *Explaining phenomena scientifically* and *Using scientific evidence*. In the component scientific knowledge a distinction is made between knowledge *of* science and knowledge *about* science. Knowledge of science is categorized in four types of systems: Physical systems, Living systems, Earth and space systems and Technology systems.

Table 1.1 Characteristics of science teaching in general and vocational lower and upper secondary education

	General secondary	Vocational secondary
Upper secondary	Grades 10–11 or 12	Grades 9–10
	All students take science course ANW (knowledge about science), mainly in grade 10	Only few streams include science/ technology courses
	About 50% of students in science streams	No attention for knowledge about science
Lower secondary	Grades 7–9	Grades 7–8
	Science for all students	Science for all students
	Science in context	Science in context
	Knowledge of science	Knowledge of science
	Mainly physical, living and technology systems	Mainly physical, living and technology systems
	Includes scientific competencies	Less attention for scientific competencies
	Regular use of graphs and diagrams	Regular use of graphs and diagrams
	Mainly open test questions	Mainly multiple choice questions

Knowledge about science is categorized in the nature of scientific inquiry and scientific explanations. The attitude component refers to *Interest in science* and *Support for scientific inquiry*.

1.2.2 Comparison of Dutch Science Education with the PISA Framework

When we compare this PISA Framework with current trends in Dutch science education we should distinguish between lower and upper secondary education and between general and vocational education (Table 1.1). Around the age of 12, Dutch students are selected for the different study programmes of secondary education: general secondary education prepares students for colleges and universities; vocational secondary education prepares students for further professional courses. For vocational secondary education the international grades 7 and 8 are defined as lower secondary education and international grades 9 and 10 as upper secondary education; for general secondary education international grades 7, 8 and 9 are defined as lower secondary education and grades 10, 11 and 12 (for the programme preparing for colleges grade 11 is the highest grade) as upper secondary education.

In lower secondary education, science is often taught in contexts. The contexts have a variety of purposes: to illustrate theories, to motivate students, to understand the environment, to act as a backbone of a teaching unit or to illustrate the importance of scientific knowledge. Graphs and diagrams are often used in lower secondary education, in line with educational practice in Dutch primary education. One of the first Dutch curriculum projects in which contexts were used extensively is PLON (physics), dated in the 1970s and early 1980s of the last century (Eijkelhof & Kortland, 1988; Kortland, 2005; Lijnse, Kortland, Eijkelhof, Van Genderen, & Hooymayers, 1990). Teaching science in contexts is also encouraged by more recent government guidelines for teaching, focusing on fields such as health, environment and safety.

In the knowledge domains, three systems are part of the national curriculum; *Physical systems*, *Living systems* and *Technology systems*. Only *Earth and space systems* are not addressed in the science curriculum for lower secondary education. Knowledge *about* science is usually not emphasized in lower secondary education, especially not in vocational education. Scientific competencies get more attention in general than in vocational secondary education.

In the field of the attitude component, no guidelines exist: it is left to the teachers to raise interest in science. The environmental aspect is addressed in two core objectives:

30. The pupil learns that humans, animals and plants are interrelated with each other and their environment, and that technological and physical applications may influence both positively and negatively the sustainable quality of the environment.
35. The pupil learns about care and learns to care for himself, for others and for his environment, and learns how to positively influence his own safety and that of others in different living situations (living, learning, working, going out, traffic).

In conclusion, many components of the PISA Scientific Literacy Framework should be familiar to 15-year-old Dutch students who have (almost) completed lower secondary education.

In upper secondary education teaching in contexts is less common, but not absent. Currently, curriculum development projects are carried out in which a stronger link between concepts and contexts is promoted. It is expected that the new curricula will be implemented from 2013 onwards (Pilot & Bulte, 2006).

Dutch 15-year-old students in general secondary education in 2006 attended an obligatory course on science for public understanding in grade 10, called ANW (de Vos & Reiding, 1999; Eijkelhof & Kapteijn, 2000). In this course the emphasis is on knowledge *about* science. Topics deal with the nature and history of scientific knowledge, its social and ethical implications and the relation between science and technology.

The conclusion is that Dutch students are fairly familiar with science in contexts. This may explain the results found by Nentwig, Rönnebeck, Schöps, Rumann, and Carstensen (2009), who found that Dutch students perform comparatively higher on "high-context" items than on "low-context" science-items.

As contexts are partly used in science classes to motivate students and contexts are often used in Dutch science teaching, one may expect a positive attitude towards science amongst Dutch students. Such an expectation is also reasonable as during the last two decades much effort has been put into activities that were intended to increase students' interest in science, for instance by large government sponsored programmes such as Axis and Platform Bèta-Techniek.

1.2.3 Expected Strengths and Weaknesses of Dutch Students

Aims of this study are to answer the following research questions:

1. Which strengths and weaknesses do Dutch students show in their PISA results, compared to students in countries with similar economic, cultural and geographic characteristics?
2. Which differences in PISA-results could be found between Dutch students in general and vocational education?
3. What are the attitudes of Dutch students towards science compared to students in countries with similar economic, cultural and geographic characteristics?
4. In which way could the results be explained by curricular developments in Dutch science education?

In view of the current curricula for Dutch secondary science education one may hypothesize the following strengths and weaknesses of Dutch students:

1. Relatively low results of Dutch students in items on *Earth and Space systems*;
2. Relatively better results in scientific competencies and the area *Knowledge about science* for students in general secondary education, compared to students in vocational education;
3. Relatively positive results of Dutch students in items that ask for interpretation of graphs and diagrams;
4. Relatively positive results of Dutch students in attitudes towards science.

1.3 Methods

1.3.1 Introduction

Beyond the global ranking, PISA reports performance on various subscales. From a curriculum innovation perspective, it is worthwhile to search for specific strengths and weaknesses at the level of subscales. However, an analysis at the subscale level still only reveals differences with regard to these predefined categories. Therefore, we considered an analysis at the level of individual items worthwhile to reveal new and unexpected patterns: the work by Nentwig et al. (2009) that was mentioned above provides an example of this.

The present analysis was done both to search for differences between Dutch students and those from a relevant peer group of seven neighbouring countries and to search for differences between students in Dutch vocational and general secondary education.

For the cross-country comparison, PISA science-items were selected for which Dutch students performed relatively better or worse than students from

Table 1.2 Performance of students from selected countries on the science scale

Country	Science score	OECD rank
Finland	563	1
Netherlands	525	6
Germany	516	8
United Kingdom	515	9
Belgium	510	13
Sweden	503	16
Denmark	496	18
Norway	487	24

seven neighbouring countries. Table 1.2 shows the general science results for the selected countries: Belgium, Denmark, Finland, Germany, Norway, Sweden and the United Kingdom. These countries were chosen for reasons of roughly similar economic, cultural and geographic characteristics in comparison with the Netherlands.

1.3.2 Method of Analysis at the Item Level

Some countries perform better than others and some items are more difficult than others. For instance, it is well known that on average Finnish students perform better than students from other countries; one may therefore expect that in general the items are easier for them than for students from the other countries. However, those differences are not the focus of our study: this chapter deals with deviant score patterns at the item level. For this purpose, for each item we consider the difference between the item's p-value for an individual country and the item's p-value across all eight countries. These differences per item were standardized into z-scores by subtracting the average difference in p-values between the Netherlands and the average for all eight countries for all science-items (on average the difference between p-values for the Netherlands and across all countries is 2.0) from the difference for the individual item and dividing this by the standard deviation of all differences (SD = 5.9). In this way the results for various countries have been compared so that we are able to show which items are relatively easier ($z \geq +1.0$) or more difficult ($z \leq -1.0$) for Dutch students in comparison with other countries; the difference between the p-value across the eight countries and the p-value of the Netherlands is one standard deviation or more larger than the average difference. In the same way, science-items were identified on which students in vocational secondary education programmes scored relatively lower than students in general secondary education programmes.

1.4 Results

1.4.1 Strengths and Weaknesses of Dutch Students

1.4.1.1 Relatively Difficult Items

Several items appeared to be relatively difficult for Dutch students. Of these items only one has been released. This item starts with a newspaper article about the history of inoculation (for the full item see the Appendix). Dutch students score remarkably low on Question 4:

Question 4: MARY MONTAGU S477Q04

Give one reason why it is recommended that young children and old people, in particular, should be vaccinated against influenza (flu).

This item belongs to the category *Living systems* and requires the competence Explaining phenomena scientifically. The item appeared to be much more difficult for Dutch students than for their peers in the other countries (z-score = −2.2). Students from Denmark were also less successful (z-score = −1.1), while students from Germany, Sweden and Norway had less problems with this item (z-scores resp. 1.3, 1.0 and 1.7). In order to answer this item it is not necessary to read the newspaper article, but question 4 itself demands careful reading.

Six relatively difficult items are multiple response items in which two or three sub items have to be answered with 'Yes' or 'No'. One open difficult item uses an agricultural authentic context which may not be familiar to Dutch students living mainly in urbanized areas. Two items require size classification of three objects in the area of Earth and space systems. The result might be explained by the fact that Earth and space systems are not commonly dealt with in lower secondary science education.

1.4.1.2 Relatively Easy Items

Also from the set of relatively easy items for Dutch students only one has been released. Question 5 of this item on the greenhouse effect (the complete item is included in the Appendix) has a remarkably positive result for Dutch students. After a discussion about the correlation between temperature and CO2 level, the question is:

Question 5: GREENHOUSE S114Q05

André persists in his conclusion that the average temperature rise of the Earth's atmosphere is caused by the increase in the carbon dioxide emission. But Jeanne thinks that his conclusion is premature. She says: "Before accepting this conclusion you must be sure that other factors that could influence the greenhouse effect are constant". Name one of the factors that Jeanne means.

This item belongs to the category *Earth and space systems* and requires the competence *Explaining phenomena scientifically*. The item requires students to look beyond the data presented in two graphs which show a correlation between the average earth temperature and the annual emission of carbon dioxide. Students are expected to take into account other factors than the ones presented, to explain the rise of temperature. This item is classified as one of the most difficult items in the PISA-test. The average p-value for Dutch students is 0.34 and for students from the eight countries (including the Netherlands) 0.22. Dutch students perform better on this item than their peers in other countries (z-score = 1.7). This result might be explained by the fact that the greenhouse topic is often dealt with in Dutch science courses. On average students from the other countries hardly differed in their answers on this item. Two relatively easy (unreleased) items require the interpretation of one or two graphs. This result might be explained by the fair amount of time spent on reading graphs and diagrams in Dutch primary and lower secondary education.

1.4.2 Differences Between Students in General and Vocational Secondary Education

Dutch students in general secondary education perform far better on the PISA-scale and subscales than their peers in vocational secondary education. This result might be expected as students are placed in programmes for secondary education according to their academic abilities around the age of 12. However, the differences were much larger for some items than for others. For 22 items the results of vocation students were remarkably low and for 18 items they were fairly good.

Weak results of this group of students were found for open-constructed response items; 13 of these items were analysed by comparing answers of 20 randomly selected students from each school-type. Almost all students made an effort to answer the open-constructed response items; no significant difference was found in efforts to answer these items. Across the questions we detected two types of problems for vocational students: (1) using the correct terms, and (2) describing all relevant steps in processes to explain phenomena.

Items in the area of *Knowledge about science* and *Using scientific evidence* appeared to be most difficult for vocational students, as was expected as the course on public understanding of science (ANW) is not part of their study programme.

The relative easy questions for vocational students were mainly in a multiple choice format. This might be explained by various factors: (1) no need to use their own wording, which they find difficult, (2) familiarity with multiple choice questions in school, and (3) the guessing factor: a difficult multiple choice question might show higher results than a difficult open question in which guessing is not rewarded.

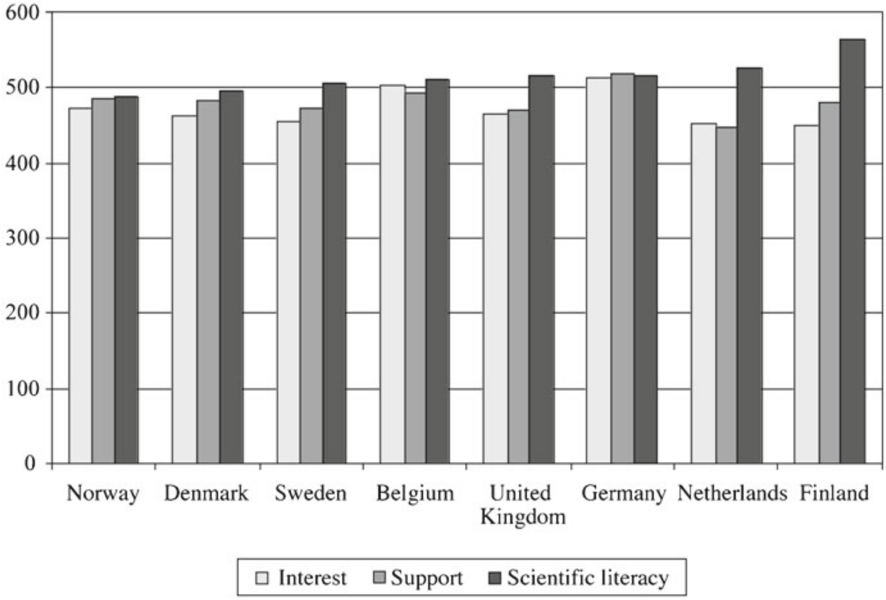


Fig. 1.1 Average score for scientific literacy, interest in and support for science per country

1.4.3 Analyses of the Attitudinal Scales

According to the PISA-ranking for science (OECD, 2007), Dutch students perform rather well on science literacy (rank 9 worldwide); but compared to other countries they show little interest in science and offer little support for scientific inquiry. Comparison of the average scores for the selected countries on scientific literacy, and interest in and support for science shows an inverse relation between scientific literacy and attitudes towards science (Fig. 1.1). However, within most countries (including the Netherlands) correlations between scientific literacy and attitudes towards science tend to be positive.

For the Netherlands, the negative attitudes towards science might be partly explained by the low self-concept for science of Dutch students (Fig. 1.2). The Dutch students' self-concept might be relatively low, because they compare themselves with their classmates, who are high-performing in comparison with their peers in many other countries. Self-concept is a strong predictor for both interest in science (stat = 24.1; SE = 1.00; p < .00) and support for scientific inquiry (stat = 25.9; SE = 1.13; p < .00) of Dutch students and might therefore explain the low scores of Dutch students on the attitude-scales.

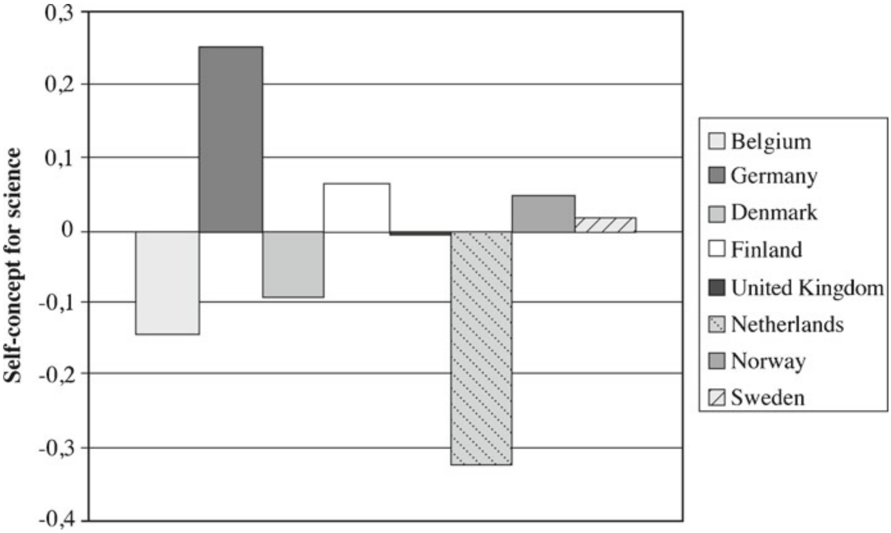


Fig. 1.2 Average score on the scale for 'self-concept in science' per country

1.5 Conclusions

From the comparison between the PISA 2006 Scientific Literacy Framework and current developments in Dutch science education one may conclude that most components of this framework are familiar to Dutch 15-year-old students. This regards especially the use of contexts in science education. The field *of Knowledge of science* is well covered by the lower secondary science curricula, with the exception of the area *Earth and Space systems*. *Knowledge about science* is familiar to most students in general secondary education but not to students in vocational education.

The framework is also in line with curriculum development trends in upper secondary science education in which strong emphasis is put on a relation between concepts and contexts. Therefore it is no surprise that Dutch students do relatively well on highly contextualized PISA-items (Nentwig et al., 2009).

Compared to their peers in other countries Dutch students do quite well on items which require the interpretation of graphs. However, not on all aspects Dutch students score high on the PISA-test. Multiple response items are relatively difficult for them, probably because this type of items is not common in Dutch education. Specific weaknesses in *Earth and Space systems* questions were not detected, although we would have expected this on the basis of Dutch science curricula for secondary education.

Dutch students in vocational education are relatively less successful in answering open-constructed response items than their peers in general education. Especially describing more complicated processes appears to be difficult for them. Dutch students in vocational programmes also have difficulties with items on knowledge *about* science and on the use of scientific evidence. This is not surprising in view of the content and nature of science education in vocational secondary education in the Netherlands.

Dutch students score low on attitude items as do students in other countries with the highest scientific literacy scores. However, within countries the correlation between attitude and scientific literacy is usually positive. In the Netherlands the low self-concept for science might be an explanation for the negative attitudes towards science.

1.5.1 Implications for Dutch Science Education

The PISA science results so far have not had large effects on Dutch policy in science education, such as reported in some other countries (Dolin & Krogh, 2010; Grek, 2009; Takayama, 2008). This is understandable as the results are relatively good compared to neighbouring countries and because the PISA approach is in line with trends in Dutch science education. Furthermore, PISA is not meant to evaluate science curricula and content-specific curriculum recommendations per country would require a different design of the study. However, this does not mean that the PISA results should be disregarded.

In Dutch vocational education students appear to be weak in open-constructed response items. Some people might accept that as a consequence of lower academic abilities of these students, competences on which they have been selected at the age of 12. We would argue that this should not mean that teachers in vocational study programmes should avoid activities in which students are required to argue scientifically. These students are also citizens of a society in which competences as these are required in order to participate. In our view the PISA Framework should apply to all students. More effort should be put in developing activities which require scientific arguing and are in accordance with the interest and potential of students in vocational secondary education. It would require some more emphasis on open-constructed response items. Use might be made of the results of investigations on discourse in science classrooms (Driver, Newton, & Osborne, 2000; Kelly, 2007; von Aufschnaiter, Erduran, Osborne, & Simon, 2008).

The PISA Framework includes the field of *Knowledge about science*. This field is given specific attention in the ANW course on public understanding of science in general upper secondary education, mainly taught in grade 10, the first year of upper secondary education. Recently the conditions for teaching this course have been weakened: the course is now only obligatory for students in the study programme preparing for university and less lesson-time is available for ANW than before. In the future this might have an effect on the students' level of knowledge *about* science. Care should be taken that this kind of valuable learning outcomes are not disregarded.

Science education should not only prepare students for living in modern society but also for post-secondary education in the area of science and technology. This means that Dutch students should not only be competent in highly contextualized items but many of them should also be able to answer less contextualized and abstract items, since in higher education courses are usually more abstract and less contextualized.

Finally, the most worrying results of PISA science are that Dutch students show relatively low interest in, and support for science. This negative attitude is surprising in view of many efforts to make science more attractive in both formal and informal education. Many activities in this area have been initiated in recent years by universities, colleges, science centres and the media. It might be that the effects of all these efforts have yet to be materialized in the attitudes of students in the period after 2006. Some signs of this are visible in the recent increase in the number of students opting for science streams in upper secondary education.

Further investigations are necessary to understand the nature of this attitude and to explore possibilities to realize a more positive attitude towards science.

Appendix: PISA items

Read the texts and answer the questions that follow.

The Greenhouse Effect: Fact or Fiction?

Living things need energy to survive. The energy that sustains life on the Earth comes from the Sun, which radiates energy into space because it is so hot. A tiny proportion of this energy reaches the Earth.

The Earth's atmosphere acts like a protective blanket over the surface of our planet, preventing the variations in temperature that would exist in an airless world.

Most of the radiated energy coming from the Sun passes through the Earth's atmosphere. The Earth absorbs some of this energy, and some is reflected back from the Earth's surface. Part of this reflected energy is absorbed by the atmosphere.

As a result of this the average temperature above the Earth's surface is higher than it would be if there were no atmosphere. The Earth's atmosphere has the same effect as a greenhouse, hence the term greenhouse effect.

The greenhouse effect is said to have become more pronounced during the twentieth century.

It is a fact that the average temperature of the Earth's atmosphere has increased. In newspapers and periodicals the increased carbon dioxide emission is often stated as the main source of the temperature rise in the twentieth century.

A student named André becomes interested in the possible relationship between the average temperature of the Earth's atmosphere and the carbon dioxide emission on the Earth. In a library he comes across the following two graphs.

André concludes from these two graphs that it is certain that the increase in the average temperature of the Earth's atmosphere is due to the increase in the carbon dioxide emission.

Question 3: GREENHOUSE S114Q03

What is it about the graphs that supports André's conclusion?

Question 4: GREENHOUSE S114Q04

Another student, Jeanne, disagrees with André's conclusion. She compares the two graphs and says that some parts of the graphs do not support his conclusion.

Give an example of a part of the graphs that does not support André's conclusion. Explain your answer.

Question 5: GREENHOUSE S114Q05

André persists in his conclusion that the average temperature rise of the Earth's atmosphere is caused by the increase in the carbon dioxide emission. But Jeanne thinks that his conclusion is premature. She says: "Before accepting this conclusion you must be sure that other factors that could influence the greenhouse effect are constant".

Name one of the factors that Jeanne means.

Mary Montagu

Read the following newspaper article and answer the questions that follow.

The History of Vaccination

Mary Montagu was a beautiful woman. She survived an attack of smallpox in 1715 but she was left covered with scars. While living in Turkey in 1717, she observed a method called inoculation that was commonly used there. This treatment involved scratching a weak type of smallpox virus into the skin of healthy young people who then became sick, but in most cases only with a mild form of the disease.

Mary Montagu was so convinced of the safety of these inoculations that she allowed her son and daughter to be inoculated.

In 1796, Edward Jenner used inoculations of a related disease, cowpox, to produce antibodies against smallpox. Compared with the inoculation of smallpox, this treatment had less side effects and the treated person could not infect others. The treatment became known as vaccination.

Question 2: MARY MONTAGU S477Q02

What kinds of diseases can people be vaccinated against?

A Inherited diseases like haemophilia.
B Diseases that are caused by viruses, like polio.
C Diseases from the malfunctioning of the body, like diabetes.
D Any sort of disease that has no cure.

Question 3: MARY MONTAGU S477Q03

If animals or humans become sick with an infectious bacterial disease and then recover, the type of bacteria that caused the disease does not usually make them sick again.

What is the reason for this?

A The body has killed all bacteria that may cause the same kind of disease.
B The body has made antibodies that kill this type of bacteria before they multiply.
C The red blood cells kill all bacteria that may cause the same kind of disease.
D The red blood cells capture and get rid of this type of bacteria from the body.

Question 4: MARY MONTAGU S477Q04

Give one reason why it is recommended that young children and old people, in particular, should be vaccinated against influenza (flu).

References

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching, 46*(8), 865–883.

de Vos, W., & Reiding, J. (1999). Public understanding of science as a separate subject in secondary schools in the Netherlands. *International Journal of Science Education, 21*, 711–719.

Dolin, J., & Krogh, L. B. (2010). The relevance and consequences of PISA science in a Danish context. *International Journal of Science and Mathematics Education, 8*(3), 565–592.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287–312.

Eijkelhof, H. M. C., & Kapteijn, M. (2000). Algemene Natuurwetenschappen (ANW): A new course on public understanding of science for senior general secondary education in the Netherlands. In R. T. Cross & P. J. Fensham (Eds.), *Science and the citizen. For educators and the public* (pp. 189–199). Melbourne: Arena Publications, special issue of Melbourne Studies in Education.

Eijkelhof, H. M. C., & Kortland, K. (1988). Broadening the aims of physics education. In P. J. Fensham (Ed.), *Development and dilemmas in science education* (pp. 282–305). London: Falmer Press.

Grek, S. (2009). Governing by numbers: the PISA 'effect' in Europe. *Journal of Educational Policy, 24*, 23–37.

Kelly, G. J. (2007). Discourse in science classrooms. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research in science education* (pp. 443–469). Mahwah, NJ: Lawrence Erlbaum.

Kortland, J. (2005). Physics in personal, social and scientific contexts – A retrospective view on the Dutch physics curriculum development project PLON. In P. Nentwig & D. Waddington (Eds.), *Making it relevant. Context based learning of science* (pp. 67–89). Münster: Waxmann.

Lijnse, P. L., Kortland, K., Eijkelhof, H. M. C., Van Genderen, D., & Hooymayers, H. P. (1990). A thematic physics curriculum: A balance between contradictory curriculum forces. *Science Education, 74*(1), 95–103.

Nentwig, P., Rönnebeck, S., Schöps, K., Rumann, S., & Carstensen, C. (2009). Performance and levels of contextualization in a selection of OECD countries in PISA 2006. *Journal of Research in Science Teaching, 46*(8), 897–908.

OECD. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Paris: OECD.

OECD. (2007). *PISA 2006. Science competencies for Tomorrow's world. Vol. 1 – Analysis*. Paris: OECD.

Pilot, A., & Bulte, A. M. W. (2006). The use of 'contexts' as a challenge for the chemistry curriculum: its successes and the need for further development and understanding. *International Journal of Science Education, 28*, 1087–1112.

Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education, 44*(4), 387–407.

von Aufschnaiter, C., Erduran, S., Osborne, J., & Simon, S. (2008). Arguing to learn and learning to argue: Case studies of how students' argumentation relates to their scientific knowledge. *Journal of Research in Science Teaching, 45*(1), 101–131.

Chapter 2
Using Mathematical Competencies to Predict Item Difficulty in PISA: A MEG Study

Ross Turner, John Dossey, Werner Blum, and Mogens Niss

Abstract This paper reports an analysis of features of mathematics assessment items developed for the OECD's Programme for International Student Assessment survey (PISA) in relation to a set of six mathematical competencies. These competencies have underpinned the PISA mathematics framework since the inception of the PISA survey; they have been used to drive mathematics curriculum and assessment review and reform in several countries; and the results of the study are therefore likely to be of interest to the broad mathematics education community.

We present a scheme used to describe this set of mathematical competencies, to quantify the extent to which solution of each assessment item calls for the activation of those competencies, and to investigate how the demand for activation of those competencies relates to the difficulty of the items. We find that the scheme can be used effectively, and that ratings of items according to their demand for activation of the competencies are highly predictive of the difficulty of the items.

Keywords Mathematical competencies • PISA • Assessment • Item characteristics

R. Turner, M.Sc., DipEdPsych (✉)
Australian Council for Educational Research (ACER), 19 Prospect Hill Road,
Private Bag 55, Camberwell, VIC 3124, Australia
e-mail: turner@acer.edu.au

J. Dossey, Ph.D.
Department of Mathematics (Emeritus), Illinois State University,
65111 East Crystal Ridge, Tucson, AZ 85739, USA
e-mail: jdossey@ilstu.edu

W. Blum
Mathematics Education, Institute for Mathematics, University of Kassel,
D-34109 Kassel, Germany
e-mail: blum@mathematik.uni-kassel.de

M. Niss
Mathematics and Mathematics Education, IMFUFA/NSM, Roskilde University,
P.O. Box 260, DK 4000 Roskilde, Denmark
e-mail: mn@ruc.dk

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_2,

2.1 Introduction

What are the factors that influence the difficulty of PISA mathematics survey items? The publication of data from the PISA 2003 survey (OECD, 2004), when mathematics was the major survey domain, has enabled a deep study of cognitive factors that influence the difficulty of mathematics items. The framework on which that survey was based (OECD, 2003) outlines a set of mathematical competencies originally described in the work of Mogens Niss and his Danish colleagues (see Niss, 2003; Niss & Hoejgaard, 2011). Such an understanding of item difficulty has the potential to guide the construction of new items to better assess the full range of the PISA mathematics scale, as well as to enhance the reporting of student performance associated with PISA assessments.

To the extent that these "Niss competencies" have resonance in various national curricula (e.g. in Denmark; see the official guidelines from the Ministry of Education: www.ug.dk/uddannelser/professionsbacheloruddannelse/enkeltfag), have been used to evaluate curriculum outcomes and even have acted as drivers of curriculum and assessment reform (e.g. in Germany; see Blum, Drueke-Noe, Hartung, & Köller, 2006, and in Catalonia, Spain; see Planas, 2010), an understanding of their influence on the difficulty of mathematics items will have far wider relevance than just within the PISA context, and will contribute more generally to an important area of knowledge in mathematics education.

The authors led an investigation that has extended over several years, beginning in October 2003. They built on earlier work aiming at understanding student achievement in mathematics developed by de Lange (1987), Niss (1999), and Neubrand et al. (2001). The investigation has focused on six mathematical competencies which are a re-configuration of the set of competencies which have been at the heart of the Mathematics Framework for PISA from the beginning (see OECD, 2003, 2006). These competencies describe the essential activities when solving mathematical problems and are regarded as necessary prerequisites for students to successfully engage in "making sense" of situations where mathematics might add to understanding and solutions. These six competencies were:

- Reasoning and argumentation
- Communication
- Modelling
- Representation
- Solving problems mathematically (referred to as Problem solving)
- Using symbolic, formal and technical language and operations (referred to as Symbols and formalism).

These competencies are not meant to be sharply disjoint. Rather, they overlap to a certain degree, and mostly they have to be activated jointly in the process of solving mathematical problems.

The initial investigation has consisted in developing operational definitions of these six competencies, and in describing four levels of demand for activation of each competency (see Sect. 2.2). PISA survey items have been analysed in relation to those definitions and level descriptions, by the application of a set of rating values to each item for each competency. The resulting ratings have then been analysed as predictor variables in a regression on the empirical difficulty of the items, derived from the PISA 2003 survey data. The item ratings have been found to be highly predictive of the difficulty of the items (see Sect. 2.3). In addition, statistical studies have been conducted examining other variables, such as the four PISA mathematical content strands (quantity, space and shape, change and relationships, and uncertainty), the PISA contexts in which the item items are presented to students (personal, education/occupational, public, scientific, and intra-mathematical), as well as the item formats themselves (various forms and combinations of multiple-choice, closed constructed-response, and open constructed-response items). None of these studies showed that these variables, acting singly or in combination with one another, explained significant proportions of the variation observed in item difficulty.

In this paper, we will present those competency definitions and level descriptions as well as the essential outcomes of the analysis conducted.

2.2 The Competency Related Variables

The material following in Table 2.1 contains the definitions and difficulty level descriptions of the six mathematical competencies used in this investigation so far. Each of the six competencies has an operational definition bounding what constitutes the competency as it might appear in PISA mathematics assessment items and then four described levels (labeled as levels 0, 1, 2, and 3) of each variable.

2.3 Analysis of the Application of the MEG Item Difficulty Framework

The following analyses provide an examination of the efficacy of the MEG Item Difficulty Framework in explaining the variability present in student performance on the 48 items common to the PISA 2003 and PISA 2006 mathematics assessments. We examine this efficacy from a number of perspectives: correlation of variable code average values, coder consistency, percentage of variance explained, consistency across assessments, and factor structure.

2.3.1 Psychometric Quality

2.3.1.1 Correlation of Variable Average Code Values

Table 2.2 contains the results of a correlation of the coding data associated with each of the six competency-based variables. Note that in this and subsequent tables, the competency labels are abbreviated as follows: REA for Reasoning and

Table 2.1 MEG item-difficulty coding framework

Symbols and formalism	
Variable-definition	**Symbols and formalism** [Understanding, **manipulating**, and **making use** of symbolic expressions within a mathematical context (including arithmetic expressions and operations), governed by mathematical **conventions and rules**; understanding and **utilising constructs** based on definitions, rules and **formal systems**.]
Level 0	No mathematical rules or symbolic expressions need to be activated beyond fundamental arithmetic calculations, operating with small or easily tractable numbers.
Level 1	Make direct use of a simple functional relationship, either implicit or explicit (for example, familiar linear relationships); use formal mathematical symbols (for example, by direct substitution or sustained arithmetic calculations involving fractions and decimals) or activate and directly use a formal mathematical definition, convention or symbolic concept.
Level 2	Explicit use and manipulation of symbols (for example, by algebraically rearranging a formula); activate and use mathematical rules, definitions, conventions, procedures or formulae using a combination of multiple relationships or symbolic concepts.
Level 3	Multi-step application of formal mathematical procedures; working flexibly with functional or involved algebraic relationships; using both mathematical technique and knowledge to produce results.
Reasoning and Argumentation	
Variable-definition	**Reasoning and argumentation** [Logically rooted thought processes that explore and link problem elements so as to **make inferences** from them, or to **check a justification that is given** or **provide a justification** of statements.]
Level 0	Make direct inferences from the instructions given.
Level 1	Reflect to join information to make inferences, (for example to link separate components present in the problem, or to use direct reasoning within one aspect of the problem).
Level 2	Analyse information (for example to connect several variables) to follow or create a multi-step argument; reason from linked information sources.
Level 3	Synthesise and evaluate, use or create chains of reasoning to justify inferences or to make generalisations, drawing on and combining multiple elements of information in a sustained and directed way.
Problem solving	
Variable-definition	**Solving problems mathematically** [Selecting or devising, as well as implementing, a mathematical strategy to solve problems arising from the task or context.]
Level 0	Take direct actions, where the strategy needed is stated or obvious.
Level 1	Decide on a suitable strategy that uses the relevant given information to reach a conclusion.
Level 2	Construct a strategy to transform given information to reach a conclusion.
Level 3	Construct an elaborated strategy to find an exhaustive solution or a generalised conclusion; evaluate or compare strategies.

(continued)

Table 2.1 (continued)

Modelling	
Variable-definition	**Modelling** [**Mathematising** an extra-mathematical situation (which includes structuring, idealising, making assumptions, building a model), or **making use** of a given or constructed model by **interpreting** or validating it in relation to the context.]
Level 0	Either the situation is purely intra-mathematical, or the relationship between the real situation and the model is not needed in solving the problem.
Level 1	Interpret and infer directly from a given model; translate directly from a situation into mathematics (for example, structure and conceptualise the situation in a relevant way, identify and select relevant variables, collect relevant measurements, make diagrams).
Level 2	Modify or use a given model to satisfy changed conditions or interpret inferred relationships; or choose a familiar model within limited and clearly articulated constraints; or create a model where the required variables, relationships and constraints are explicit and clear.
Level 3	Create a model in a situation where the assumptions, variables, relationships and constraints are to be identified or defined, and check that the model satisfies the requirements of the task; evaluate or compare models.
Communication	
Variable-definition	**Communication** [Decoding and **interpreting** statements, questions and tasks; including **imagining** the situation presented so as to **make sense** of the information provided; **presenting and explaining** one's work or reasoning.]
Level 0	Understand a short sentence or phrase relating to a single familiar concept that gives immediate access to the context, where it is clear what information is relevant, and where the order of information matches the required steps of thought.
Level 1	Identify and extract relevant information. Use links or connections within the text that are needed to understand the context and task, or cycle within the text or between the text and other related representation/s. Any constructive communication required is simple, but beyond the presentation of a single numeric result.
Level 2	Use repeated cycling to understand instructions and decode the elements of the context or task; interpret conditional statements or instructions containing diverse elements; or actively communicate a constructed description or explanation.
Level 3	Create an economical, clear, coherent and complete description or explanation of a solution, process or argument; interpret complex logical relations involving multiple ideas and connections.
Representation	
Variable-definition	**Representation** [**Interpreting**, translating between, and **making use** of given representations; **selecting** or **devising** representations to capture the situation or to present one's work. The representations referred to are depictions of mathematical objects or relationships, which include equations, formulae, graphs, tables, diagrams, pictures, textual descriptions, concrete materials]

(continued)

Table 2.1 (continued)

Level 0	Directly handle a given representation, for example going directly from text to numbers, reading a value directly from a graph or table, where minimal interpretation is required in relation to the situation.
Level 1	Select and interpret one standard or familiar representation in relation to a situation.
Level 2	Translate between or use two or more different representations in relation to a situation, including modifying a representation; or devise a simple representation of a situation.
Level 3	Understand and use a non-standard representation that requires substantial decoding and interpretation; or devise a representation that captures the key aspects of a complex situation; or compare or evaluate representations.

Table 2.2 Correlations of competency-based variable values based on the coding of 48 PISA mathematics items by eight coders

	AVGSYM	AVGREA	AVGPS	AVGMOD	AVGCOM
AVGREA	0.283				
	0.051				
AVGPS	0.301*	0.721*			
	0.038	0.000			
AVGMOD	0.606*	0.455*	0.401*		
	0.000	0.001	0.005		
AVGCOM	0.405*	0.471*	0.100	0.267	
	0.004	0.001	0.497	0.066	
AVGREP	0.062	0.314*	0.303*	0.261	0.082
	0.676	0.030	0.036	0.073	0.581

Cell contents: correlation, *p*-value

*Correlation significantly different from 0 at the 0.05 level

Argument; PS for Problem Solving; MOD for Modelling; COM for Communication; REP for Representation; SYM for Symbols and Formalism. The prefix 'AVG' indicates the average code value across the eight coders on the relevant competency.

2.3.1.2 Coder Consistency

Coder consistency can be approached from two perspectives. The first is the degree to which coders' actual coding of the items correlated with codings they had initially given the items in another coding of the same items 2-years previously. This would be an analysis of intra-coder consistency. The other examination of coder consistency would be an examination of the degree to which the eight coders tended to code in common for a given item relative to the competencies. Such consistency would be an example of inter-coder consistency.

Table 2.3 Intra-coder consistency for common PISA 2003/PISA 2006 items

Coder\competency	SYM	REA	PS	MOD	COM	REP
Coder 1	0.804*	0.803*	0.805*	0.885*	0.847*	0.860*
Coder 2	0.644*	0.906*	0.777*	0.856*	0.855*	0.884*
Coder 3	0.505*	0.575*	0.459*	0.380*	0.652*	0.703*
Coder 4	0.369*	0.428*	0.438*	0.579*	0.462*	0.404*

*r is significantly different from 0 at the 0.05 level

Intra-coder consistency. Within-coder data only exist for four of the eight coders in our sample. In addition, there have been minor changes in the description of the competency-related codes for some of the variables that may have slightly altered the use of the codes during the intervening 2 years. These cautions notwithstanding, correlations were conducted for the six competency-related variables for each of the coders for whom there was complete data for the two separate codings of the 48 items. The results are in Table 2.3 Note that all of the observed correlations were significantly different from 0 at the 0.05 level.

There was an interesting pattern in the intra-coder correlations of the coders' work. The coders are numbered in ascending order from most consistent to least consistent. This ordering also matches the ordering of amount of experience and coding the four coders had with using the MEG Item Difficulty Framework. This suggests, perhaps, that coders become more consistent with increased familiarity with the framework and its use, and that training in the use of the framework will be an important issue for the future.

Inter-coder consistency. A second approach to coder consistency lies in examining the degree to which the eight coders actually give the same code to an item for a given competency. In essence, this is asking to what degree the eight individual MEG coders give the same numerical code to an item for any one of the six competency-related variables. This analysis can be approached from a variety of perspectives. Historically, most researchers have been satisfied with finding the Pearson product moment correlation of the coders over the set of items related to a given competency area. More recently, researchers dealing with content coding and curricular studies have shifted toward the use of Cronbach's α along with more emphasis on individual item and coder patterns of behaviour (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shrout & Fleiss, 1979; von Eye & Mun, 2005).

The data showing the codes each of the eight individual coders awarded for each item have been collected and analysed to determine the consistency of the eight coders for each of the six competency-based variables. Table 2.4 contains a variety of information points for each item. In addition to Cronbach's α value for each competency-based variable, data are provided showing the distribution of ranges between minimum and maximum codes given to individual items for the competency variable in the coding. Note that a range of 3 for an individual item indicates that it was coded as being at both Level 0 and Level 3 by different coders.

The examination of the values of Cronbach's α for the six competency-based variables shows considerable consistency with the exception of the Reasoning

Table 2.4 Inter-coder consistency data for the six competency-based variables

	SYM	REA	PS	MOD	COM	REP
Cronbach's α	0.89	0.62	0.90	0.81	0.95	0.83
Average code	1.42	1.52	1.54	1.48	1.58	1.40
Code range = 0	3	0	1	1	0	3
Code range = 1	26	25	24	24	24	27
Code range = 2	15	21	19	22	20	14
Code range = 3	4	2	4	1	4	4
Coding outliers	C5 = 2 C8 = 2	C2 = 1 C5 = 1	C8 = 4	C6 = 1	C1 = 1 C4 = 1 C5 = 1 C8 = 1	C5 = 2 C8 = 2

and argumentation variable. An examination of the individual item and response data in general did not immediately indicate a reason for the lower consistency value observed.

The next row in the table indicates the average coding value given across each of the 48 items in each of the competency-based variables' actual coding for this study.

The remaining rows of the table provide a great deal of information about how the 48 items were coded within each of the competency-based variables. The term "Code Range" in the left hand column refers to the range of codes, which is the value of maximum code awarded minus the value of the minimum code. For example a code range of 0 would indicate that all eight coders agreed on the code awarded an item. A code range of 1 would indicate that all coders were awarding one of two adjacent codes. A code range of 3, however, would indicate that at least one coder had awarded a code of 0 while another coder had awarded a code of 3 to an item. Items for which this occurred were flagged for extra analysis. The coders were numbered C1–C8 and the final row in the table indicates which coders were "outliers" in the coding of the individual item receiving a code range of 3.

2.3.2 *Results of Difficulty Analyses*

2.3.2.1 Predicting Variance Explained

The degree to which the six competency related variables add to the explanation of variance in the item difficulty scores associated with student performance for the PISA 2003 and PISA 2006 mathematics surveys was analysed using first the "best subsets" approach, and then through a separate multivariate regression analysis of the data.

Best Subset Regressions: Analysis of the PISA 2003 data, the implementation of the "best subset" regression approach, which is sometimes called the "all possible regressions" approach, resulted in the information shown in Table 2.5 (Chatterjee & Price, 1977; Draper & Smith, 1966).

Table 2.5 Best subset analysis of the PISA 2003 item logit values using the average competency variable scores as predictors

Vars	*R*-Sq	Adj. *R*-Sq	C-p	s	AVGSYM	AVGREA	AVGPS	AVGMOD	AVGCOM	AVGREP
1	48.1	47.0	38.6	0.94742		X				
2	68.8	67.4	7.7	0.74293	X	X				
3	72.4	70.5	4.0	0.70669	X	X	X			
4	73.8	71.3	3.7	0.69643	X	X	X		X	
5	74.0	70.9	5.4	0.70206	X	X	X	X	X	
6	74.2	70.5	7.0	0.70705	X	X	X	X	X	X

The values in each row show data associated with various sets of predictor variables and the percentage of variance in the PISA 2003 item difficulty variability they predict. This percentage is given by the value in the Adjusted R-square column. The data show the best predictor variables as the number of variables used in the mode goes from 1 to 6.

An examination of the table provides a number of observations. First, the percent of variance predicted (Adjusted R-Squared value) increases up to a four-variable model and then decreases slightly thereafter for the best models with five or six predictor variables. The degree to which the increased value of prediction increases will be discussed later.

It is interesting that the one best competency predictor is the Reasoning and argumentation variable. The entrance of additional predictor variables in building best models with more variables show the entry order of Symbols and formalism, Problem solving, Communication, Modelling, and Representation. The latter two variables do not appear to add to the explanatory power achieved using only the first four.

Table 2.6 shows the same analysis conducted using the PISA 2006 item difficulty estimates. The results are very similar in that the four-variable model appears the best in numerical value and the first four variables entering are the same: Reasoning and argumentation, Symbols and formalism, Problem solving, and Communication. However, there is a slight difference in the order of the entrance of the remaining two variables into the predictor models. Here the next is Representation, followed then by Modelling. However, the data suggest that the addition of these latter two variables does not improve the prediction based on the four variables common to both the PISA 2003 and PISA 2006 data.

Overall, these best subset regression analyses indicate that the four variables of Reasoning and argumentation, Symbols and formalism, Problem solving, and Communication provide the best structure for maximizing the prediction of item difficulty in PISA as defined by item logit values. Additional analysis of the relative contributions of each of these will appear in the next analyses.

Multiple regressions: Table 2.7 contains the results of a stepwise regression employing all possible competency-based variables for the prediction of the PISA 2003 item difficulty logit values. The algorithm was structured to select the best single predictor, and then add the next best single predictor that would add a significant amount of explanatory power. This process iterates, adding variables to the regression equation until the point when the addition of any other variable to the regression equation would no longer make a statistically significant increase in the amount of item difficulty variance explained.

This regression equation indicates that the three competency-based variables, in order of explanatory power are Reasoning and argumentation, Symbols and formalism, and Problem solving. This model predicts 70.5% of the variability in the PISA item difficulty data, when the R-squared value is adjusted. While the addition of the variable Communication would have pushed the R-squared value to 71.8, the gain would not have been statistically significant over the variance explained by this three-variable model.

Table 2.6 Best subset analysis of the PISA 2006 item logit values using the average competency variable scores as predictors

Vars	*R*-Sq	Adj. *R*-Sq	C-p	s	AVGSYM	AVGREA	AVGPS	AVGMOD	AVGCOM	AVGREP
1	48.5	47.4	40.5	0.93106		X				
2	69.3	68.0	8.4	0.72646	X	X				
3	73.2	71.4	4.0	0.68674	X	X	X			
4	74.8	72.4	3.4	0.67423	X	X	X		X	
5	75.0	72.0	5.1	0.67938	X	X	X	X	X	
6	75.0	71.4	7.0	0.68676	X	X	X	X	X	X

Table 2.7 Stepwise regression for the explanation of variability in the PISA 2003 item difficulty logit values

Step	1	2	3
Constant	−2.212	−2.524	−2.573
AVGREA	1.64	1.32	0.87
T-value	6.53	6.44	3.20
AVGSYM		1.09	1.02
T-value		5.46	5.33
AVGPS			0.67
T-value			2.39
S	0.947	0.743	0.707
R-Sq	48.09	68.78	72.38

The resulting regression equation is:
PISA 2003 = −2.573 + 0.87 * AVGREA + 1.02 * AVGSYM + 0.67 * AVGPS

Table 2.8 Stepwise regression for the explanation of variability in the PISA 2006 item difficulty logit values

Step	1	2	3
Constant	−2.212	−2.521	−2.572
AVGREA	1.62	1.31	0.85
T-value	6.59	6.53	3.22
AVGSYM		1.08	1.01
T-value		5.53	5.42
AVGPS			0.68
T-value			2.52
S	0.931	0.726	0.687
R-Sq	48.53	69.35	73.22

The resulting regression equation is: PISA 2003 = −2.572 + 0.85 * AVGREA + 1.01 * AVGSYM + 0.68 * AVGPS

Carrying out the same stepwise regression approach using the PISA 2006 item difficulty logit data as the dependent variable, we obtain the results shown in Table 2.8. As in the case of the PISA 2003 data, the same three variables, Reasoning and argumentation, Symbols and formalism, and Problem solving enter in the same order. In this case, the three variables explain 71.4% of the variability in the item difficulty values when the adjusted R-squared value is computed.

A comparison of the coefficients show that there is no difference between the models developed from the PISA 2003 and PISA 2006 data. In like manner, there is no difference in the ascending order in which the three statistically significant predictor variables enter into the equations. In both cases, the calculation of the Durbin-Watson statistic and other residual diagnostics indicate that these models are sound and free of common biasing factors sometimes found in regression model building.

2.3.2.2 Factor Analysis

A factor analysis was conducted to examine the structure of the space spanned by the six competency-based variables. A principal components factor analysis of the

Table 2.9 Factor analysis of the competency-related variable codings

Sorted unrotated factor loadings and communalities

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
AVGREA	–0.833	–0.229	–0.355	–0.213	–0.129	0.258
AVGMOD	–0.762	0.181	0.450	0.064	–0.418	–0.075
AVGPS	–0.736	–0.430	0.020	–0.441	0.188	–0.207
AVGSYM	–0.666	0.538	0.353	0.022	0.361	0.110
AVGCOM	–0.554	0.484	–0.605	0.267	–0.002	–0.148
AVGREP	–0.438	–0.588	0.088	0.665	0.112	0.008
Variance	2.758	1.140	0.826	0.758	0.360	0.149
%Var	0.460	0.190	0.138	0.126	0.062	0.025

correlation matrix of the competency-based variable scores for the 48 items revealed the findings shown in Table 2.9. An examination of the data indicates that there were two factors having eigenvalues greater than one. Given that each variable contributes a value of 1 to the eigenvalues total, only the factors having eigenvalues in the end greater than one are considered significant and retained for further study.

An examination of the percent of variance described by the first two factors show that they account for a total of 64% of the variance in the codings. Factor 1's strongest loadings are Reasoning and argumentation, Modelling, Problem solving, and Symbols and formalism. This might be considered, given the values, a balanced factor similar to a generalised academic demand factor. Factor 2's strongest loadings are Symbols and formalism decreased by Representation and Problem solving. This second factor might be considered as describing increased item demand related to the requirement to decode and deal with Symbols and formalism and Communication in the absence of Problem solving and the demand to interpret and manipulate Representations. One might liken this to adding demand for reading and symbol manipulation as it occurs without enacting problem solving strategies or multiple representations of mathematical concepts or operations.

An important remark: It might seem that a certain subset of those six competencies will already serve all purposes and that the others are unnecessary. However, a subset of competencies proved to be sufficient only for explaining item difficulty and only in the particular case of PISA tests. In other cases, other subsets might have more explanatory power. More importantly, the competencies serve a much broader purpose than only explaining item difficulty. For the most important purpose, that is describing proper mathematical activities and thus formulating the essential aims that students ought to achieve through school mathematics, all competencies are indispensable.

2.4 Present Status of the Study

The foregoing data provide sufficiently strong evidence of the role played by the mathematical competencies, as defined in Table 2.1, in influencing variability in item difficulty on the PISA mathematics survey items. At present, illustrations of the way the competencies play out to influence difficulty in particular items are being developed, along with an elaborated coding manual for researchers who have not been involved in the development of the MEG model. This coding manual will be central in the next stage of the study, as it will be used with researchers unfamiliar with it and the coding of PISA items, but familiar with coding structures. They will be asked to code the 48 PISA items and their results will be compared with those of the MEG members.

The planned next steps are as follows. Based on this experience and revisions that may result from observing these coders and their work, a broader field test shall be conducted where new individuals, familiar with the PISA project, will be asked to use the coding instruction manual without any other assistance to code the 48 items. Their coding results and written comments shall again be used to further the development of the model and manual for either one more round of field testing or release as a PISA technical report.

Two further developments of this study might be to investigate the extent to which the scheme could be used to predict the difficulty of newly developed PISA mathematics items; and to investigate its applicability to other (non-PISA) mathematics items.

Curriculum statements in many countries reflect the importance of the competencies on which this study has focused. It can be expected that the relationship between cognitive demand for the activation of these competencies and the empirical difficulty of the mathematical tasks that call for such activation, whether in the PISA context or in other contexts, will be of deep interest to teachers, teacher educators and others involved in mathematics education around the world.

References

Blum, W., Drueke-Noe, C., Hartung, R., & Köller, O. (2006). *Bildungsstandards Mathematik: konkret*. Berlin: Cornelsen-Scriptor.

Chatterjee, S., & Price, B. (1977). *Regression Analysis by Example*. New York: Wiley.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

De Lange, J. (1987). *Mathematics, Insight and Meaning*. Utrecht: CD-Press.

Draper, N. R., & Smith, H. (1966). *Applied Regression Analysis*. New York: Wiley.

Neubrand, M., Biehler, R., Blum, W., Cohors-Fresenborg, E., Flade, L., Knoche, N., et al. (2001). Grundlagen der Ergänzung des internationalen OECD/PISA-Mathematik-Tests in der deutschen Zusatzerhebung. *Zentralblatt für Didaktik der Mathematik, 33*(2), 45–59.

Niss, M. (1999). Kompetencer og uddannelsesbeskrivelse. *Uddannelse, 9*, 21–29.

Niss, M. (2003). Mathematical Competencies and the Learning of Mathematics: The Danish KOM Project. In A. Gagatsis & S. Papastavridis (Eds.), *3rd Mediterranean Conference on Mathematical Education* (pp. 115–124). Athens: The Hellenic Mathematical Society.

Niss, M., & Hoejgaard, T. (eds.) (2011) *Competencies and mathematical learning. Ideas and inspiration for the development of mathematics teaching and learning in Denmark.* English edition, October 2011 (IMFUFAtekst no. 485). Roskilde: Roskilde University.

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

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

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

Planas, N. (2010). *Pensar i comunicar matemàtiques*. Barcelona: Fundació Prepedagogic.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin, 86*(2), 420–428.

Von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement: manifest variable methods*. Mahwah: Lawrence Erlbaum. www.ug.dk/uddannelser/professionsbacheloruddannelse/enkeltfag. Visited 19 May 2011.

Chapter 3
PISA Mathematics in Germany: Extending the Conceptual Framework to Enable a More Differentiated Assessment

Michael Neubrand

Abstract Assessing mathematical literacy—as PISA does—claims for comprehensive views of the domain tested. Since mathematics is not a homogenous body of knowledge one needs inner structures of that domain in order to be able to interpret the data gained. There are several possibilities, e.g. to differentiate between the main content strands as geometry, algebra etc. However the German PISA options differentiated according to cognitive activities connected with mathematics. These activities contain the performance of procedures as well as conceptual thinking, in both intra- and extra-mathematical situations. This paper exhibits the basis of that framework, i.e. a model for mathematical tasks, and shows evidences and findings from that approach, as the cognitive balances of several tests, and the striking cognitive profiles we found in different parts of the country.

Keywords Cognitive profiles • Cognitive activities in mathematics • Domain specific assessment • Mathematical literacy • PISA mathematics • Test construction in mathematics

3.1 Introduction

Since its inception, PISA has aimed to measure student knowledge and skills with a special focus on "how well young adults are prepared to meet the challenges of the future" (OECD, 1999, p. 3; OECD, 2001). From its very beginnings, however—internationally, but especially in Germany—PISA also pursued a kind of meta-goal: to stimulate thinking about the objectives of the tested domains within an education system. This meta-goal was made more or less explicit, at least in the domain of

M. Neubrand (✉)
Mathematics Education, Institute for Mathematics, Carl-von-Ossietzky
University, D-26111 Oldenburg, Germany
e-mail: michael.neubrand@uni-oldenburg.de

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_3,

mathematics, where the conceptualization of the domain as “mathematical literacy” was a signal to the community of mathematics educators to restructure their thinking about how mathematics is addressed in schools, and how the outcomes of mathematics education should be evaluated.

Given the very specific situation of Germany, the international PISA test was complemented by a national extension study. The German school system is defined by three major characteristics. First, students are streamed to separate secondary school tracks from the age of 10 years on; the academic track (*Gymnasium*) leads to higher education, whereas the other tracks prepare students for vocational training and careers. Second, these tracks not only organize the system, but differ in the ways that subjects are taught. Third, although the literacy-based approach taken in PISA—which is rooted in pragmatic traditions of education and in the Freudenthal approach to mathematics pedagogy—is widely endorsed by the mathematics education community, it was clear that the reality of German classrooms in the year 2000 was far from that ideal. Germany wanted to respond to this situation by implementing national options that were compatible with the PISA framework.

Consequently, Germany took the opportunity offered by the OECD to develop an additional national option very seriously. These deliberations resulted in an additional day of testing immediately after the international PISA test. To this end, additional items were constructed on the basis of a newly developed framework (Neubrand et al., 2001), the aim of which was “to broaden and differentiate” the international PISA framework (at that time: OECD, 1999). The basic “mathematical literacy” approach was retained, but adapted and extended on the basis of the German discourse on mathematics education.

This paper (a) discusses the need for differentiated assessment categories from a general viewpoint, (b) sketches the approach taken in the German framework, and (c) outlines key findings. It argues that communicating the results of a (mathematics) achievement test in a more differentiated way is valuable, as it allows outcomes to be reported in a manner that is closer to the reality and needs of mathematics teaching and its development. (Indeed, the newly established educational standards for mathematics in Germany would surely not have been possible without PISA; the standards draw heavily on the PISA-based idea of conceptualizing mathematics achievement under more than one dimension, even in the cognitive sense we focus on here [see Blum, Drüke-Noe, Hartung, & Köller, 2006].)

3.2 The Need to Differentiate: Mathematics Achievement Is Not Homogeneous Across Countries

Inspection of the results of PISA 2000 revealed considerable between-country differences in performance on the mathematics items—even among countries at the same achievement level. Figure 3.1 presents the average percent correct data of the

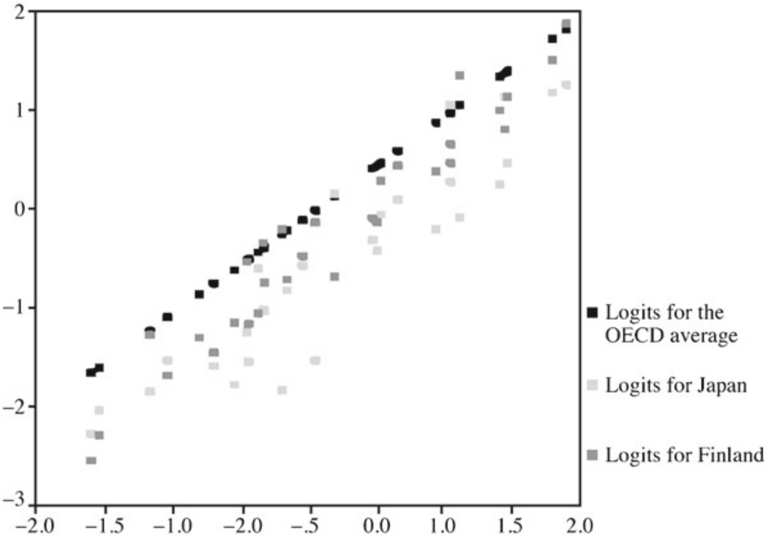


Fig. 3.1 Average percent correct data of all PISA 2000 mathematics items (transformed to logits). Horizontal axis: logits for the OECD average, vertical axis: logits for Japan (*light gray*), Finland (*dark gray*), and the OECD average (*black*)

PISA 2000 mathematics items converted into logits, with the two high performing countries Japan and Finland being plotted against the OECD average (Fig. 3.1; first presented in Neubrand & Neubrand, 2003). The OECD average thus forms the diagonal as a reference line. Each set of three dots plotted vertically above one another represents a single PISA 2000 mathematics item.

Figure 3.1 shows two things. First, it illustrates the striking differences in the achievement of the Japanese students relative to that of the students in the average of the OECD countries. Moreover, Japanese students' performance on the items shows surprisingly large variability: although, in some cases, their achievement is close to the OECD average, the rate of correct answers provided by Japanese students differs considerably depending on the item in question. In Finland, in contrast, the pattern of students' achievement is much closer to the OECD average. Moreover, as the items become more difficult, the rate of correct answers provided by Finnish students approaches the OECD average, whereas the easier items are much easier for Finnish students. Thus, Finland achieves its place in the PISA "top ten" by having—so to say—the best weak students, whereas Japan seems to achieve its position by taking very different approaches to mathematics.

Both observations—which can be substantiated by regression analyses (Neubrand & Neubrand, 2003)—point to an inner in-homogeneity of mathematics itself. Further conceptualization is thus required to provide a reasonable picture of what mathematics is about.

3.3 A Model for Mathematical Tasks

A deeper understanding of the origins of high and low achievement requires a finely graded picture of what a task is (structural approach), which features explain its difficulty (predictivity), and how these features can be composed into a system. A series of aspects, all with a background in mathematics education, therefore have to be considered: the structural aspect (What is a task?), the content aspect (What is mathematical activity?), the broadness aspect (How is it possible to gain a full and manageable picture of students' mathematics literacy using a limited set of items?), and the content validity aspect (How is it possible to retain the PISA-specific literacy approach within a broader set of items?)

The well known cycle of mathematical modeling (e.g., Blum, Galbraith, Henn, & Niss, 2007) can serve not only as a means of describing the translation of a problem situation in a real-world context into a mathematical representation, but also as an overall model of the process of solving a mathematical task (Fig. 3.2).

On this basis, a more detailed model of mathematical tasks was developed for use in the German interpretation of the PISA 2000 results (Neubrand, 2004). As its "kernel" this model incorporates the four rectangles shown in the center of Figure 3.3, namely the four structural elements defining the structure and character of any mathematical task (Neubrand, 2002, 2006), and thus the four basic decisions to be made in categorizing tasks: What kind of thinking is dominant during the working-out phase of the modeling cycle—procedural or conceptual (in the sense of Hiebert, 1986)? Is a mathematizing and/or problem solving activity needed? In other words, does a problem situation need to be translated into a mathematical representation, according to the modeling cycle? Does the working-out phase involve a single step or multiple

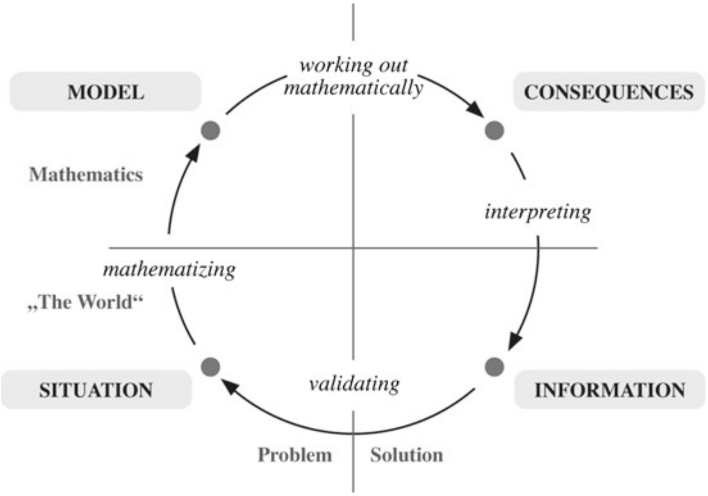


Fig. 3.2 The cycle of cognitive activities during a mathematical modeling process (see Blum et al., 2007; Klieme, Neubrand, & Lüdtke, 2001; Neubrand et al., 2001)

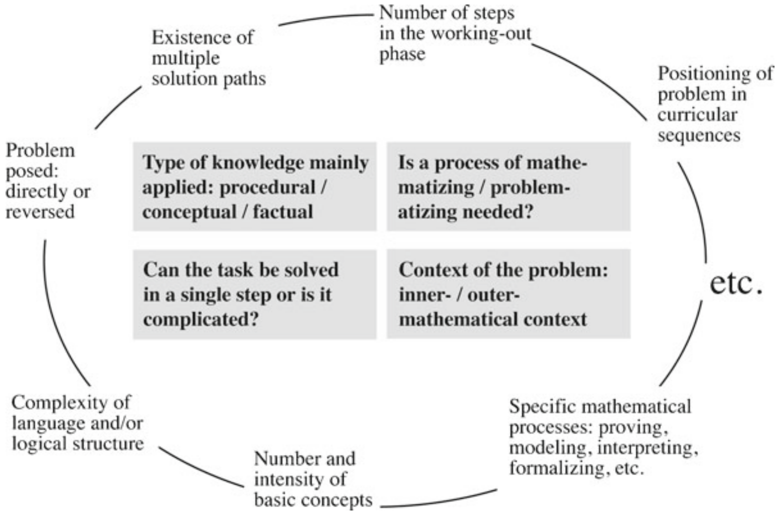

Number of steps in the working-out phase
Positioning of problem in curricular sequences
etc.
Specific mathematical processes: proving, modeling, interpreting, formalizing, etc.
Number and intensity of basic concepts
Complexity of language and/or logical structure
Problem posed: directly or reversed
Existence of multiple solution paths
Type of knowledge mainly applied: procedural / conceptual / factual
Is a process of mathematizing / problematizing needed?
Can the task be solved in a single step or is it complicated?
Context of the problem: inner- / outer-mathematical context

Fig. 3.3 The model of mathematical tasks used in the German interpretation of the PISA 2000 results (Neubrand, 2004)

steps? Is the task set in a real-world context, or is the whole process—including translation from a problem situation to a mathematical representation—done inner-mathematically?

These four dimensions qualitatively distinguish the different kinds of cognitive activities involved in the solution of mathematics tasks. They are not themselves systematically related to the difficulty of tasks. They can be further condensed to reflect the basic decisions associated with a task—or what we call the "three types of mathematical activity" (Neubrand & Neubrand, 2004):

- *Technical tasks*: Procedural thinking only; one- or multi-step; no problem solving or modeling activity needed; all activities are inner-mathematical.
- *Procedural modeling and/or problem-oriented tasks:* Mathematizing and/or problem solving is necessary; the origin can be either a real-world situation (modeling) or a mathematical situation implying inner-mathematical problem solving activities; mostly *procedural thinking* during the working-out process; one- or multi-step.
- *Conceptual modeling and/or problem-oriented tasks:* Mathematizing and/or problem solving is necessary; the origin can be either a real-world situation (modeling) or a mathematical situation implying inner-mathematical problem solving activities; mostly *conceptual* thinking during the working-out process; one- or multi-step.

As mathematics is characterized by both procedural and conceptual thinking (Hiebert, 1986; Kilpatrick, 2001), a balanced assessment of mathematical literacy can be expected to show a roughly equal distribution of tasks over at least the two

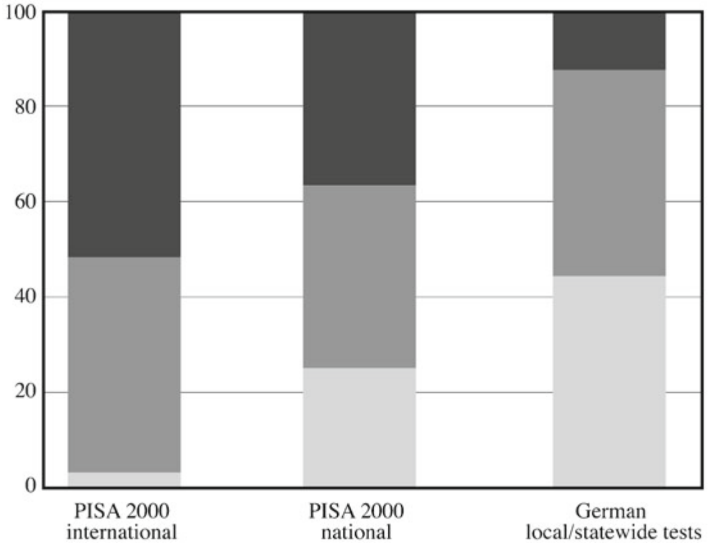


Fig. 3.4 Characteristics of assessments, according to the three types of mathematical activities (in % of items administered). Legend: "Technical tasks": *light gray*. "Procedural modeling tasks": *gray*. "Conceptual modeling tasks": *dark gray*

classes of modeling/problem solving tasks. As Fig. 3.4 shows, however, application of this structural distinction to different assessments reveals marked differences in their composition.

Specifically, the international PISA assessment achieved a balance between tasks requiring procedural and conceptual thinking, as did the German national PISA option, which also included a selection of "technical" tasks as part of its extended conceptualization of mathematical literacy. However, some local or statewide tests administered in Germany at the same time as PISA did not achieve such a balance. Technical activities played the dominant role in these assessments.[1]

3.4 Features of Mathematical Tasks

The classification of tasks into the three types of mathematical activities does not suffice as a model for tasks. The features of mathematical tasks shown around the "kernel" as forming the "periphery" of Fig. 3.3 represent several further

[1] The overemphasis of technical tasks is increasingly emerging (see Neubrand, 2002, for TIMSS) to be a characteristic of German mathematics classes Analyses conducted in the context of a representative study of mathematics teachers' professional knowledge, the COACTIV study (Baumert et al., 2010), revealed that up to 90% of the tasks set in high-stakes classroom tests are of the technical type (Jordan et al., 2008).

aspects associated with tasks from various special viewpoints. All features in this open list have been shown to specifically influence the difficulty of PISA tasks. Moreover—and providing convincing evidence for the value of differentiated assessment—which features of a task are relevant for predicting its difficulty has been shown to depend on which of the three types of mathematical activities is considered (Neubrand, Klieme, Lüdtke, & Neubrand, 2002). The detailed, in-depth analyses of the PISA 2000 items (both the international items and the items of the German national option) conducted by the German PISA mathematics expert group (Neubrand, 2004) revealed various such cases. Three of them are described in the following:

First, Neubrand et al. (2002) conducted regression analyses to establish how various item features are related to the difficulty of those items. They showed (see also Neubrand & Neubrand, 2004) that some frequently examined task features, such as the number of steps required to solve an item, indicate the difficulty of only two item classes: the technical and the procedural modeling items. In contrast, the difficulty of the conceptual modeling/problem solving items was found to depend on the general demands of the modeling and/or problem solving process, as reflected in the establishment of proficiency levels in the PISA mathematics framework (OECD, 1999).

Second, Cohors-Fresenborg, Sjuts, and Sommer (2004) found that the complexity of the language of a task serves as an indicator of its difficulty, and noted that the ability to deal with formulas helps students to solve the respective tasks. Third, Blum, vom Hofe, Jordan, and Kleine (2004) found that the intensity with which tasks elicit basic mathematical concepts ("Grundvorstellungen" in the sense of vom Hofe, Kleine, Blum, & Pekrun, 2005) is one of the most decisive factors in predicting the difficulty of a task, but only for the modeling items.

3.5 Profiles of Mathematical Achievement

A further step forward was the identification of "profiles" of mathematical achievement in several populations in terms of the three types of mathematical activity. The German education system is characterized by a heterogeneous structure across the 16 federal states (*Bundesländer*). The states not only have different curricula, they also differ in terms of school structures, distributions of students to the various school types, and final examinations. Consequently, the traditions and methods of mathematics teaching also differ across states. Accordingly, the differences found in the achievement scores of students across the 16 states are of high political interest. From the viewpoint of mathematics education, however, the more interesting question is whether different patterns or profiles of achievement can be detected—in other words, whether certain states show characteristic strengths and weaknesses. This is indeed the case. Plotting the individual states' achievement by the three types of mathematical activities reveals profiles that can be traced back to specific curricular decisions in the states. One striking effect revealed by this differentiated

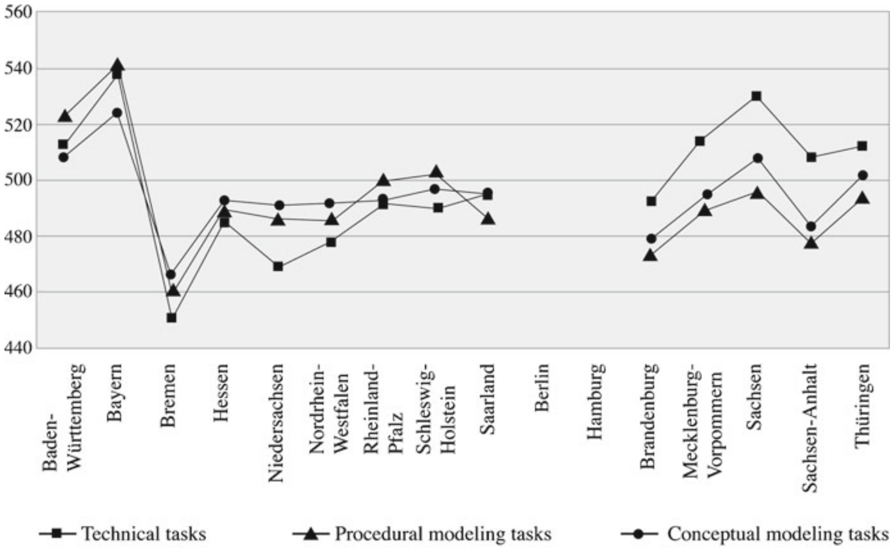


Fig. 3.5 Profiles of mathematical achievement in the 16 German federal states by the three types of mathematical activities: data from PISA 2000 (Neubrand & Neubrand, 2004). Berlin and Hamburg are not included in the figure as these city states did not meet the PISA sampling requirements

analysis is the emphasis on technical performance in the former East German states, which are shown on the right side of Fig. 3.5. In contrast, most of the former West German States (on the left side of the diagram) showed their weakest performance on the technical tasks (Fig. 3.5).[2]

3.6 Advantages of Differentiated Assessment

Why is differentiated assessment so important for the further development of testing in mathematics education? There are three answers to this question.

First, the differentiated characterization of test items according to a theory-based set of features highlights the key characteristics of a test (from a mathematics education perspective). This information is crucial in test construction, as it allows the construction of a fairly balanced test (e.g., to test mathematical literacy). As the three types of mathematical activities mirror the inner structure of mathematics from a

[2] As a similar pattern of findings emerged for some analogous sub-competencies in the PISA science test (Rost, Carstensen, Bieber, Prenzel, & Neubrand, 2003), these data can usefully inform discussion of curricula and their implementation. Note that boys' and girls' performance on the three types of mathematical activities also differed (Neubrand & Neubrand, 2004).

cognitive perspective, this approach assures broadness of test construction, and thus allows "technical" competencies to be incorporated within a literacy perspective.

Second, differentiated assessment provides insights into issues pertaining to the development and reform of mathematics teaching. Such issues arise when, for example, one asks whether TIMSS and PISA really test the "same" mathematical achievement, or whether a country's improved performance on a test can be explained by certain curricular decisions. For example, the regression analysis conducted by Neubrand and Neubrand (2003) revealed that much of the difference between Germany and Japan is attributable to the fact that Japanese students are considerably better able to cope with geometrical drawings than are German students. Data of this kind can inform content-related pedagogical decisions. Similarly, reforms can be targeted more precisely if the data show an emphasis on certain ways of knowing and teaching: Such data make it easier to identify shortcomings—and potentially even trace those shortcomings back to certain didactic traditions (which may then be questioned).

Third, as countries show different achievement on the item level (see Fig. 3.1), it can be concluded that classroom practices must differ, and it may claimed that some practices are more conducive to certain aspects of mathematics achievement. However, as differentiated assessment also reveals different assessment behaviors across countries—as shown by the comparison between Japan and Finland in Sect. 3.1, it is important to be aware that there are several ways to succeed in PISA (as in any test). Accordingly, results must always be interpreted against the background of a country's didactic practices.

Differentiated assessment thus underlines that, as already noted in the context of TIMSS Video, "mathematics teaching is a cultural activity" (Stigler & Hiebert, 1999). This statement evidently applies not only to teaching methods, but also to the content-based characteristics of students' mathematical achievement.

References

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., et al. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180.

Blum, W., Drüke-Noe, C., Hartung, R., & Köller, O. (Eds.). (2006). *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen [Educational standards for mathematics at lower secondary level: Sample tasks, suggestions for lessons, ideas for continuing education]*. Berlin, Germany: Cornelsen.

Blum, W., Galbraith, P. L., Henn, H.-W., & Niss, M. (Eds.). (2007). *Modelling and applications in mathematics education: The 14th ICMI study*. Berlin, Germany: Springer.

Blum, W., vom Hofe, R., Jordan, A., & Kleine, M. (2004). Grundvorstellungen als aufgabenanalytisches und diagnostisches Instrument bei PISA [Basic mathematical concepts (‚Grundvorstellungen') as a diagnostic and analytic instrument in PISA]. In M. Neubrand (Ed.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland: Vertiefende Analysen im Rahmen von PISA 2000* (pp. 145–158). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Cohors-Fresenborg, E., Sjuts, J., & Sommer, N. (2004). Komplexität von Denkvorgängen und Formalisierung von Wissen [Complexity of thought processes and formalization of knowledge]. In M. Neubrand (Ed.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland: Vertiefende Analysen im Rahmen von PISA 2000* (pp. 109–144). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Hiebert, J. (Ed.). (1986). *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ: Lawrence Erlbaum.

Jordan, A., Krauss, S., Löwen, K., Blum, W., Neubrand, M., Brunner, M., et al. (2008). Aufgaben im COACTIV-Projekt: Zeugnisse des kognitiven Aktivierungspotentials im deutschen Mathematikunterricht [Tasks in the COACTIV project: Evidence of the potential for cognitive activation in German mathematics instruction]. *Journal für Mathematik-Didaktik, 29*(2), 83–107.

Kilpatrick, J. (2001). Understanding mathematical literacy: The contribution of research. *Educational Studies in Mathematics, 47*, 101–116.

Klieme, E., Neubrand, M., & Lüdtke, O. (2001). Mathematische Grundbildung: Testkonzeption und Ergebnisse [Mathematical literacy: Test conception and results]. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 139–190). Opladen, Germany: Leske & Budrich.

Neubrand, J. (2002). *Eine Klassifikation mathematischer Aufgaben zur Analyse von Unterrichtssituationen: Schülerarbeitsphasen und Selbsttätigkeit in den Stunden der TIMSS-Video-Studie [A classification of mathematics tasks for the analysis of instructional situations: Phases of independent student work in the TIMSS Video Study lessons].* Berlin/Hildesheim, Germany: Franzbecker.

Neubrand, M. (Ed.). (2004). *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland: Vertiefende Analysen im Rahmen von PISA 2000 [Mathematical competencies of students in Germany: In-depth analyses in the context of PISA 2000].* Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Neubrand, J. (2006). The TIMSS 1995 and 1999 video studies. In F. K. S. Leung, K.-D. Graf, & F. J. Lopez-Real (Eds.), *Mathematics education in different cultural traditions: A comparative study of East Asia and the West: The 13th ICMI Study* (pp. 291–318). Berlin, Germany: Springer.

Neubrand, M., Biehler, R., Blum, W., Cohors-Fresenborg, E., Flade, L., Knoche, N., et al. (2001). Grundlagen der Ergänzung des internationalen PISA-Mathematik-Tests in der deutschen Zusatzerhebung [Framework for the supplementation of the international PISA mathematics test in the German national PISA extension study]. *Zentralblatt für Didaktik der Mathematik, 33*(2), 33–45.

Neubrand, M., Klieme, E., Lüdtke, O., & Neubrand, J. (2002). Kompetenzstufen und Schwierigkeitsmodelle für den PISA-Test zur mathematischen Grundbildung [Proficiency levels and difficulty models for the PISA test of mathematical literacy]. *Unterrichtswissenschaft, 30*(2), 100–119.

Neubrand, J., & Neubrand, M. (2003). *Profiles of mathematical achievement in the PISA 2000 mathematics test and the different structure of achievement in Japan and Germany.* Paper presented at the AERA 2003 Annual Meeting, Chicago.

Neubrand, J., & Neubrand, M. (2004). Innere Strukturen mathematischer Leistung im PISA 2000 Test [Internal structures of mathematics achievement in the PISA 2000 test]. In M. Neubrand (Ed.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland: Vertiefende Analysen im Rahmen von PISA 2000* (pp. 87–108). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

OECD. (1999). *Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematical and scientific literacy*. Paris, France: OECD.

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000.* Paris, France: OECD.

Rost, J., Carstensen, C., Bieber, G., Prenzel, M., & Neubrand, M. (2003). Naturwissenschaftliche Teilkompetenzen im Ländervergleich [Aspects of science literacy in German cross-state comparison]. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann, & M. Weiß (Eds.), *PISA 2000: Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 109–128). Opladen, Germany: Leske & Budrich.

Stigler, J., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: Free Press.

vom Hofe, R., Kleine, M., Blum, W., & Pekrun, R. (2005). On the role of "Grundvorstellungen" for the development of mathematical literacy: First results of the longitudinal study PALMA. *Mediterranean Journal for Research in Mathematics Education, 4*, 67–84.

Part II
Methodological Research

Introduction: Methodological Research in Large-Scale International Assessments

Mark Wilson

During the early 1980s, Torsten Husén, one of the founders of the IEA, came to the campus[1] where I was enjoying my doctoral training, and visited another one of those founders, Benjamin Bloom. Bloom, being wise in the ways of academe, invited him to talk to the assembled graduate students, and Husén proceeded to explain what the founders had been thinking about when they put the IEA enterprise together. The story was that they had had two major intentions: one was to set up a situation where the different countries of the world could be seen as, in effect, a range of "natural experiments", each varying somewhat in their "treatments" of their students, but with the educational achievement surveys, and associated measures, forming a relatively comparable set of outcomes, which could be used to "test" the successes (and failures) of those unplanned experiments. A second intention was to find a mechanism to upgrade the quality of the educational researchers in those "underdeveloped" countries that had not gained the benefits of having educational research scholars trained in advanced methods. Although he did not mention it as a major aim, the tradition of large-scale surveys of educational attainment that they began has also had a profound influence on another aspect of educational research, that is, on the methodologies that are applied to educational programs, both across and within nations. The large-scale surveys, now including the continuing series of IEA-sponsored surveys, as well as the PISA surveys which are the focus of this volume, have served as an engine and a workbench for the development of methodological innovations ever since. As a young researcher from one of those "underdeveloped" countries (Australia), I was inspired by both the challenges that I saw in these early surveys, as well as the tremendous excitement to see those challenges met through innovative methods and creative applications of research methodologies.

[1] The University of Chicago.

M. Wilson, B.Sc.(Hons.), DipEd, M.Ed, Ph.D.
BEAR Center, Graduate School of Education, University of California, Berkeley,
3659 Tolman Hall 1670, Berkeley, CA 92720-1670, USA
e-mail: MarkW@berkeley.edu

These early IEA studies introduced many innovations to educational research, including the "design effect", as a way to capture the implications of non-random sampling designs on analyses of educational samples that were structured by the realities of education in the different countries (e.g., Peaker, 1974). Later innovations from within IEA and from other large-scale surveys included systematic evaluations of the characteristics of different curricula across countries (e.g., Schmidt, 1992), and the inclusion of demographic covariates in analyses and the reporting of "plausible values" (Mislevy, Johnson, & Muraki, 1992) to enhance the possibilities of secondary analysts to incorporate measurement errors (so beloved of the psychometricians) into their results. With the advent of the PISA surveys in 2000, further methodological innovations were introduced. These included: the use of multidimensional item response models to enhance estimation of item parameters, the introduction of "booklet effects" to attempt to account for disturbances brought about by item-placement patterns on parameter estimation, and the positing of a "described variable" as a means of specifying the substantive meaning of the latent variables being measured, as well as providing a proxy for a "common curriculum" across the countries (Adams, Wilson, & Wang, 1997; OECD, 2002). Later years have seen extensive introduction of computerized assessments in specific topics (and these are scheduled to become dominant in 2015).

The chapters in this section of the current volume represent a selection of the best and most innovative of the current forefront of research in this area, which continues this tradition of scholarship. The chapters are a fascinating profile of the challenges and efforts that most occupy educational researchers today. The first chapter, by the Williamses (Chap. 4 by Williams & Williams, this volume), is a classic style of innovation, bringing a methodology that has been established in other areas of scholarship (in this case, in econometrics) into the arena of large-scale educational surveys. When analyzing the survey results within and across countries, one common type of effect that one wants to investigate are what are termed "reciprocal effects" between achievement variables and other social measures—that is, where the two sorts of variables are seen to mutually influence one another. This has proven problematical for large-scale surveys in education, as it has been thought that, because they are cross-sectional in nature, the effects are confounded. However, this chapter explains and exemplifies a method, based on the "instrumental variables" technique that allows this disentanglement under certain conditions, hence opening the way to the incorporation of research and hypotheses involving reciprocal effects into the domain of educational surveys. The second chapter (Chap. 5 by Solano-Flores, Contreras-Niao, & Backhoff, this volume) tackles another perennial challenge in international surveys, the need to translate items across different languages (and cultures) and still to have confidence that this has not brought about important (and uncontrolled) changes in the meaning of the variables being measured. The chapter outlines a new and comprehensive methodology that promises to put the study of test translation on a new and sounder footing. The third chapter (Chap. 6 by Rust, Krawchuck, & Monseur, this volume) surveys the strategies employed by the PISA surveys in 2003 to address issues of non-response at the school student and item level, and shows how they were improved for the 2006 surveys. This plots just one step along a line of continual improvement in survey design and implementation that has been a hallmark of the PISA project since its inception. The fourth chapter (Chap. 7 by Frey, Seitz, & Kroehne, this volume) represents an exploration of PISA's future, as it investigates, through an ingenious simulation strategy, one of the potential benefits of computerized test administration—the possibility of adapting the selection of items delivered to a student according to that student's current estimate of ability. This offers the possibility of decreasing the number of items each student needs to take, hence setting up potential gains in either (a) reducing the time-demands of the PISA testing, or (b) increasing the number of latent variables that can be measured by any given survey. Taken as a set, these chapters reveal a lively and expert response to the current state of PISA and represent a portrait of the state of the art for methodological research on large-scale surveys.

When reading these chapters, one perspective that is hard to resist is to speculate about what lies next, just over the horizon, in methodological research for large-scale surveys. Several possibilities come readily to mind. First, the advent of computerization, already the context for one of the chapters described above, offers several different possibilities that will raise new methodological challenges. One of these possibilities is that the computer is having an effect on the underlying educational systems, and leading to the incorporation of new educational variables as targets of the surveys. For example, educational aspects of social networking are now being seen as educational achievements in their own right—a possibility that would have been seen as laughable not more than a decade ago. Yet, one prominent effort in research in educational measurement is focused on exactly that variable, though it is referred to under a different name (Wilson, 2010). A second effect will be on the complexity of the items themselves—this goes way beyond the possibility of the computerized adaptive testing described above, and raises a panoply of possibilities for new and complex assessments of processes and higher intellectual functions that we are only just now contemplating. A third effect will be on such stratagems as the "booklet effect" mentioned above—in a radically computerized assessment environment it will be a challenge merely to conceptualize what a "booklet" would constitute, let alone estimate its effect.

Second, there is a distinct possibility that the tests used for large-scale surveys could become the sources for tests used by school districts and even schools for various sorts of monitoring and evaluation. This brings about new challenges, such as the incorporation of the effects of student clustering in educational institutions like schools and school districts. This has been a potential effect all along, but it has been dissipated by the focus of the analyses on the large-scale (i.e., country and state effects). The use of a new generation of multilevel models will be required for this. Equally, use as a monitoring tool at, say, a school district level, raises the possibility that PISA could be used in a longitudinal data gathering mode, with individual students (or schools) being tracked across years. Over the years, there have been longitudinal surveys in many contexts, so this is not an area that will require new models, but the incorporation of that dimension of complexity into the PISA design will require much work and creativity, not the least of which will be to envisage the "designed variables" as spanning multiple years of schooling.

Third, one area that PISA has been an innovator in, the use of "described variables" may itself be the subject of innovation. In several subject areas, a novel development has been the postulation of "learning progressions" as a way to organize assessments. According to a recent survey:

> *Learning progressions are descriptions of the successively more sophisticated ways of thinking about an important domain of knowledge and practice that can follow one another as children learn about and investigate a topic over a broad span of time. They are crucially dependent on instructional practices if they are to occur.* (Center for Continuous Instructional Improvement [CCII], 2009)

Ways to actualize these curriculum structures in assessment terms are currently the subject of some research interest (Wilson, 2009, 2012), and may provide an important step forward, allowing the closer connection between classroom assessments and large-scale surveys that many educators and policy-makers see as highly desirable.

In conclusion, one can see from the evidence in these chapters that methodological research in and on PISA is in a very healthy state. It has a respectable place in the longer history of methodological research on large-scale international educational surveys. And there are clear ways forward that will encourage and require further efforts in the area of methodological research within the context of PISA.

References

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

Center for Continuous Instructional Improvement (CCII). (2009). *Report of the CCII panel on learning progressions in science* (CPRE Research Report). New York: Columbia University.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*(2), 131–154.

OECD. (2002). *PISA 2000 technical report*. Paris: Author.

Peaker, G. F. (1974). *An empirical study of education in twenty-one countries: A technical report* (International Studies in Evaluation VIII). Stockholm, Sweden: Almqvist and Wiksell.

Schmidt, W. H. (1992). TIMSS Curriculum Analysis: Topic Trace Mapping. *Prospects, 22*(3), 326–333.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal for Research in Science Teaching, 46*(6), 716–730.

Wilson, M. (2010, May). *21st-century measurement for 21st-century skills*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Wilson, M. (2012). Responding to a challenge that learning progressions pose to measurement practice: Hypothesized links between dimensions of the outcome progression. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science*. Rotterdam, Sense Publishers.

Chapter 4
Modeling Reciprocal Determinism in PISA

Trevor Williams and Kitty Williams

Abstract Reciprocal determinism refers to the situation where the underlying dynamic of an observed relationship is one of mutual influence. Each variable influences the other in a feedback loop. This notion is invoked in PISA to explain the relationship between students' achievements and various aspects of their learning strategies, motivations, self-beliefs and preferences. But, in PISA, as in the literature as a whole, the reciprocal determinism of theory is seldom translated into an appropriate statistical model. Rather, in statistical analyses, the notion of mutual influence tends to be abandoned and the relationship is modeled simply as a one-way effect; in this case, the effect of a particular learning strategy on achievement. The most likely reason for this inconsistency is the widely-held belief that reciprocal determinism cannot be modeled with cross-sectional data. Longitudinal, repeated-measures data are considered necessary in order to estimate reciprocal determinism as cross-lagged effects. However, it is possible to model reciprocal effects with cross-sectional data by developing nonrecursive structural equation models in which these effects are represented as a feedback loop. This approach is not without its difficulties but, to the extent that these can be resolved, analyses in which the theoretical and statistical models are consistent become possible.

The discussion below is designed to illustrate this approach using as an example a nonrecursive structural equation model in which the mutual influence of self-efficacy and performance in mathematics is represented as a feedback loop. This model is estimated in each of 33 nations using PISA 2003 data.

Keywords Reciprocal determinism • Self-efficacy • Achievement • Mathematics

T. Williams, Ph.D. (✉) • K. Williams, Ph.D.
Westat, 1600 Research Blvd, Rockville, MD 20877, USA
e-mail: trevor.williams9@gmail.com; kitty.williams99@gmail.com

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_4,

4.1 Reciprocal Determinism

Reciprocal determinism is a term coined by Bandura (1978, p.344) to describe the reciprocal influences of behavior, cognition and environment that are central to his theory of social cognition. Simply put, reciprocal determinism refers to those observed relationships where the underlying dynamic is one of mutual influence. That is, each variable influences the other in a feedback loop. Although the existence of reciprocal influence relationships among some PISA constructs is acknowledged, no attempt has been made to model these relationships statistically. The intent of this chapter is to consider how such reciprocal determinism can be modeled. A specific example is developed based on the acknowledged mutual influence of mathematics self-efficacy and mathematics achievement.

4.1.1 Reciprocal Determinism in PISA

In 2000, 2003 and 2006 PISA included variables designed to tap student approaches to learning with the view to examining their influence on student achievement. Four main themes were represented: cognitive/metacognitive learning strategies; motivational preferences and volition; self-related beliefs; and, learning situations and preferences (Artelt, Baumert, Julius-McElvany, & Peschar, 2003). Both theory and common sense suggest that many of the constructs in each of these four categories exist in a mutual influence relationship with achievement, influencing achievement and, in turn, being influenced by achievement. The notion is one of a dynamic process, which eventually reaches equilibrium. For example, the learning strategies that students employ will be reinforced, or attenuated, by feedback about their achievement; students who like mathematics are motivated to perform well on achievement tests and, as a result of positive feedback about this performance, come to like mathematics even more; students who are confident of their mathematical abilities tend to perform well on mathematics assessments, receive positive feedback about this performance, and feel more confident still; and, student preferences for certain kinds of learning situations will be reinforced to the extent that these learning situations lead to higher levels of achievement and the rewards that this brings.

PISA reports published over this period acknowledge this reciprocal determinism explicitly for some of the important constructs subsumed by the four themes noted above (Organization for Economic Cooperation and Development [OECD], 2001, p.119); in particular, for motivation (Artelt et al., 2003, p.15), engagement (Kirsch et al., 2002, p.128) and, later, for self-efficacy (OECD, 2007, p.134). It is reasonable to suppose that this same thinking applies to most, if not all, of the variables subsumed under student approaches to learning. However, in the analyses reported, the relationship of each measure to achievement is modeled simply as the one-way effect of the various strategies, motivations, beliefs and preferences on achievement. No allowance is made for a feedback influence from achievement. This is true as well in the broader literature relating to these constructs, particularly the voluminous literature associated with the effects of motivation and self-beliefs on student achievement (see, for example, Marsh & Craven, 2006; Pintrich, 2003; Pintrich & Maehr, 2002; Pintrich & Schunk, 2002). That is, reciprocal determinism is endorsed at the level of theory but the data are not modeled in a way that is consistent with this theory.

Almost certainly, the reason for this is the widely held belief that reciprocal determinism cannot be modeled with cross-sectional data. Causal inferences about reciprocal determinism are thought to be impossible, or at least implausible, in the absence of a measured time lag between cause and effect. In short, the view is that reciprocal determinism can only be modeled as cross-lagged effects using panel data with repeated measures. In fact, in most instances where attempts have been made to model reciprocal determinism, panel data have been used. See for example, Marsh and Craven (2006) who describe analyses modeling the reciprocal determinism of self-concept and achievement in this way. It follows that, since PISA data are cross-sectional, this view poses something of a dilemma for PISA analyses of the effects of student approaches to learning on achievement, given that theory suggests reciprocal determinism is the underlying dynamic.

However, under certain conditions it is possible to model reciprocal determinism with cross-sectional data. In fact, the history of such efforts stretches back about 50 years to the originator of path analysis (Wright, 1960). In social science applications other than economics, structural equation models containing feedback loops have been the approach of choice. Such models are known as nonrecursive structural equation models. Their formulation and estimation is addressed in most texts on structural equation modeling; see for example Berry (1984), Duncan (1975), Hayduk (1987), Heise (1975), Kline (2010), Maruyama (1998) and Mulaik (2009). Martens and Haase (2006) provide a recent example from psychology but, on the whole, models incorporating reciprocal effects are relatively rare in the literature. Nonrecursive models containing feedback loops are difficult to formulate in practice, in large part because they require that some parameters be specified a priori.

Nevertheless, to the extent that this is possible within the context of the PISA data, the relationship between student approaches to learning and achievement could be modeled statistically in a way that better represents the underlying complexity of these relationships. The following discussion is designed to indicate how this might be accomplished, and uses the reciprocal determinism of mathematics self-efficacy and mathematics achievement as a concrete example. Data from PISA 2003 are used to estimate a nonrecursive model in each of 33 countries.

4.2 Formulating a Nonrecursive Structural Equation Model

Structural equation models are classified as either recursive models or nonrecursive models. Recursive models are those that do not contain (a) reciprocal effects, (b) structural error terms correlated with the explanatory variables in any equation, and/or (c) structural error terms correlated with each other. Nonrecursive models

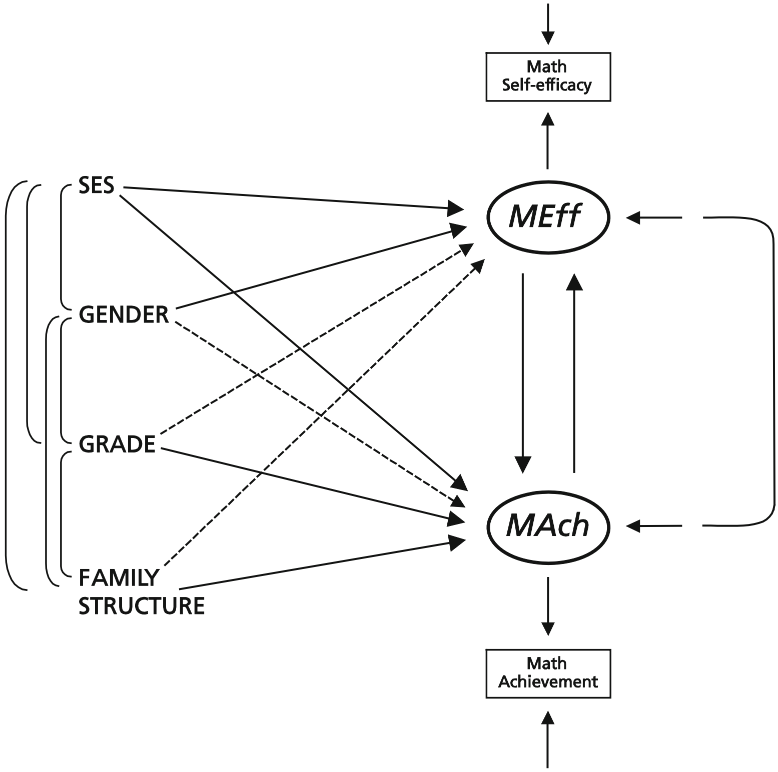


Fig. 4.1 Nonrecursive structural equation model for "Reciprocal Determinism"

may have any or all of these attributes. See Bielby and Hauser (1977, p.142), Hayduk, (1987, p.247), Heise (1975, p.153), Kaplan (2000, p.16), and Mulaik (2009, p.135). Figure 4.1 shows the nature of the nonrecursive model formulated in this instance to capture the reciprocal effects of mathematics self-efficacy and achievement.

In Fig. 4.1, consider the structural part of the model in the first instance. Mathematics self-efficacy and mathematics achievement are treated as latent variables (MEff and MAch), and each is shown as an influence on the other; self-efficacy affects achievement and, in turn, is affected by achievement in a feedback loop. As a consequence, the error terms representing the unexplained variance associated with each may not be independent, as indicated by the curved line linking these terms (Kaplan, 2000, p.16). And, ignoring the distinction between solid and dotted lines for the moment, the exogenous variables (SES, gender, grade and family structure) are seen to have effects on both self-efficacy and achievement.

The variables indicated above were all taken from the PISA data. Mathematics achievement is tapped by five plausible values (OECD, 2005b, p. 71). The measure of mathematics self-efficacy is the MATHEFF scale created by PISA (OECD, 2005a, p.291). Student socioeconomic status is measured by the index HISEI representing the highest occupational status of parents (OECD, 2005a, p.273). Gender is a student-report measure (OECD, 2005b, p.251). Family structure is a dichotomy based on a recoding of the FAMSTRUC index as follows. Families with two adults present were combined and identified as 'nuclear' families; single-parent families and all others types were combined as 'other' (OECD, 2005a, p.273). Grade represents the school year/grade for the student (taken from administrative data). In most countries students are spread across two or three grades. Dummy variables were created to capture this variation with the result that grade is represented by a single dummy variable in some (two-grade) countries, and by two dummy variables in other (three-grade) countries.

4.2.1 Identification

The identification status of a model has important implications for estimation of the model parameters and for measures of fit of the model to the data. Structural equation models may be under-identified, just-identified or over-identified. In an under-identified model there are more effect parameters to be estimated than observed variances and covariances to estimate them with. In these circumstances a unique set of parameter estimates is not possible. (Models which include a feedback loop and allow for effects from all of the exogenous variables tend to be under-identified in the first instance.) In a just-identified model the number of effect parameters and the number of observed variances and covariances are equal. Unique effect estimates are possible but measures of fit are not, since there are no degrees of freedom. An over-identified model is one in which the number of observed variances and covariances is greater than the number of parameters to be estimated and, as a result, both unique effect estimates and measures of fit are possible. Most structural equation texts provide extended discussions of identification; see, for example, Duncan (1975, p. 81), Hayduk (1987, p.143), Heise (1975, p.148), and Mulaik (2009, p. 142).

The essential challenge in the formulation of nonrecursive models is that of developing a model which is just-identified at the very least, but preferably is one that is over-identified because the latter provides for measures of fit in addition to unique parameters estimates. Under-identified models require more information in order to return a unique set of parameter estimates. There is a number of ways to accomplish this, most of which involve fixing parameters to particular values determined a priori. Fixing parameters to zero by including variables with a postulated influence on one, but not both, of the variables in the feedback loop is common. Such variables are often termed instrumental variables (Fisher, 1971). However, not just any variable will do; instrumental variables must be theoretically meaningful with plausible nontrivial effects on one of the variables in the feedback relationship, and fixed effects on the other (Duncan, 1975, p. 89). As such, they can be thought of as specific hypotheses grounded in theory and introduced into the model in order to render the model just- or over-identified. In over-identified models these fixed effects impact the fit of the model to the data. To the extent that these specifications are in error, the fit of the model will be compromised. Finding a sufficient number of plausible and effective instrumental variables in secondary analyses can be something of a challenge (Kessler & Greenberg, 1981) and probably represents the major obstacle to the development of these models.

With regard to the relationships shown in Fig. 4.1, grade, family structure and gender were treated as instrumental variables in this instance. The effect parameters fixed to zero a priori in each instance are shown by the dotted lines in Fig. 4.1. That is, the effects of grade and family structure on self-efficacy are constrained to zero, with their effects on achievement to be estimated from the data. Similarly, the effect of gender on achievement was fixed to zero leaving the effect on self-efficacy to be estimated from the data. The theoretical/substantive basis for this configuration takes the following form. Grade indexes years of exposure to learning and, as such, taps opportunity to learn. This has a direct effect on performance (Carroll, 1963). The fixed zero effect on self-efficacy implies that differences in opportunity to learn affect self-efficacy only through their effects on learning itself (achievement) and the mastery experiences it provides. Family structure (nuclear vs. other) reflects differences in social capital and economic resources between the two categories in question and these exert their influence directly through student achievement and not through self-efficacy. The zero constraint for the effect of gender on achievement is supported by the commonly accepted notion of gender stereotyping (Hyde, Fennema, Ryna, Frost, & Hopp, 2006). That is, the observed gender differences in mathematics achievement are seen to come about not because of some inherent differences in mathematical ability but because females are socialized by the expectations of significant others to have lower levels of self-efficacy which translates into lower levels of achievement. With regard to SES, the social and economic attainments of students' families were assumed to influence both self-efficacy and achievement. A more detailed development of these arguments is provided in Williams and Williams (2010).

These constraints result in a model that is over-identified. Depending on the country, there are one or two degrees of freedom available. The degrees of freedom vary as a function of whether one or two dummy variables are used to capture year/grade of schooling in a particular country. The additional information indexed by these degrees of freedom allows an examination of the fit of the model to the data.

4.2.2 Measurement Models

In addition, and independently of considerations associated with nonrecursive models, measurement models for self-efficacy and achievement were developed to allow for between-nation differences in the reliability of measurement of both constructs. Each model had the same basic structure. Using mathematics self-efficacy as an example, the construct was treated as a latent variable (MEff) with a single indicator (the self-efficacy composite score). Information on the reliability of measurement was introduced into the model by fixing the variance of the error term of this indicator to [(1 − reliability) * variance]; see Heise (1975, p.188).

4.2.3 Estimation

Once a nonrecursive model is formulated appropriately with respect to its identification status, estimation itself is straightforward for the most part and not different in principle from that for recursive models. Standard structural equation modeling software such as LISREL, MPlus, AMOS, or EQS provides for the estimation of nonrecursive models as well as the more familiar recursive models. Mplus (Muthén & Muthén, 2010) was used in this case.

Where complications arise in model estimation they arise from the design of PISA rather than the use of nonrecursive models. Two aspects of the PISA design give rise to these complications. First, since PISA adopted a Balanced Incomplete Block (BIB-spiral) design for the assessment, student scores are estimated as (five) plausible values. As a consequence, analyses involving student achievement need to be conducted five times, once for each plausible value, and the results averaged (OECD, 2005b, pp.71–80).

Second, the sampling design gives rise to two further complications. The student samples are probability samples, not simple random samples. In order to provide for national estimates of the model parameters, sampling weights need to be used in the analyses (OECD, 2005b, pp.19–30). A second complication arises out of the clustering of students that occurs in PISA samples. Standard errors estimated by the usual means are likely to be biased since these procedures assume random sampling. PISA provides for correct estimates through the use of 80 replicate weights (see OECD, 2005b, pp.31–52). (The use of a sandwich estimator may provide an acceptable alternative and one more easily implemented; see, Muthén & Muthén, 2010, p.233).

This complex sampling design also requires adjustments to the chi-square measure of fit of the model in each country. The design-effect adjustment proposed by Stapleton (2008) was applied to the chi-square values estimated in this instance.

4.3 Findings

The model described was estimated with PISA 2003 data for 33 of the 41 participating countries. Data from eight countries were excluded. In seven of the eight countries (Iceland, Japan, Korea, Norway, Poland, Sweden, and Serbia and Montenegro) more than 90% of students were in a single grade with the result that the grade variable had little variation. Data problems in the eighth country, Mexico, suggested that it should be excluded from the analyses as well (OECD, 2005a, p.243).

4.3.1 The Fit Between the Model and the Data

Since interpretations of effect estimates are warranted only if the model fits the data, the question of the fit of the model to the data in each country is taken up first. A chi-square test of fit based on the generalized likelihood ratio is available for over-identified models. The null hypothesis in this instance is that the differences between the model-implied covariance matrix and the observed covariance matrix can be attributed to sampling fluctuations alone. A significant chi-square then rejects the null hypothesis, indicates a poor fit of the model to the data and suggests that the model does not adequately reflect the 'real world' that generated the data. On the other hand, a nonsignificant chi-square means that one cannot reject the null hypothesis. Any differences between the implied and observed covariance matrices could be due to sampling fluctuations. Thus, the model and the set of estimates obtained provide a good fit to the data and have some claim to reflecting the 'real world', though one cannot dismiss the logical possibility that other models would also be consistent with same data (Hayduk, 1987, p.160).

Table 4.1 provides for each country the fit statistics obtained for the model in question, along with degrees of freedom and associated probability values. In all these statistics indicate a good fit between the model and the data in 30 of the 33 nations. In three nations (Denmark, Ireland, and Portugal) the chi-square test suggests that the model is problematic in some way, presumably as a consequence of fixing effects to zero. Further exploration of the sources of this ill-fit is possible but was not undertaken at this time. For example, one could selectively relax the constraints to improve the fit. However, since the fit was acceptable in 90% of the countries examined, the lack of fit in three countries was seen as a subsidiary issue for the present analyses, though something to be followed up at a later date. With respect to the main focus of these analyses though, a model embodying the reciprocal influence of mathematics self-efficacy and achievement, along with a common set of constrained effects, seems consistent with the data in 30 countries.

4.3.2 Parameter Estimates

Estimates of the effect parameters for the 30 nations with acceptable fit are presented in Table 4.2 as metric coefficients, coefficients expressed in their original units of measurement. Coefficients greater than +/− 1.96 times their respective standard errors are indicated with asterisks. The Low and High categories of Grade need some explanation. Grade was operationalized as one or two dummy variables depending on the grade distribution. In countries with three grades, the middle grade is omitted and there are two parameter estimates, indicating the effect of having one less, or one more, year of education than students in the middle grade. In countries where 15-year-olds are located in only two grades, the single coefficient shown indicates the effect of being in the higher of the two grades (relative to the

Table 4.1 Chi-square measures of model fit by country

	Chi-square test				Chi-square test		
Country	X^2	*df*	Probability	Country	X^2	*df*	Probability
Australia	2.012	2	0.366	Liechtenstein	2.275	2	0.321
Austria	2.006	2	0.367	Luxembourg	1.260	2	0.533
Belgium	1.839	1	0.175	Macao-China	0.701	2	0.704
Brazil	2.325	2	0.313	Netherlands	3.129	1	0.077
Canada	2.154	1	0.142	New Zealand	1.571	1	0.210
Czech Republic	2.280	1	0.131	Portugal	10.726	2	0.005
Denmark	4.850	1	0.028	Russian Federation	0.214	1	0.644
Finland	3.487	1	0.062	Slovak Republic	1.584	1	0.208
France	2.531	2	0.282	Spain	0.181	1	0.671
Germany	3.733	2	0.155	Switzerland	2.037	2	0.361
Greece	0.240	2	0.887	Thailand	0.029	1	0.865
Hong Kong-China	0.489	2	0.783	Tunisia	5.951	2	0.051
Hungary	2.156	2	0.340	Turkey	4.403	2	0.111
Indonesia	2.476	2	0.290	United Kingdom	0.092	1	0.762
Ireland	6.537	2	0.038	United States	1.454	2	0.483
Italy	1.029	1	0.310	Uruguay	2.799	2	0.247
Latvia	1.014	2	0.602				

Note: From Williams and Williams (2010). Copyright 2010 by American Psychological Association

$p>.05$ indicates acceptable degree of fit between the model and the data

lower grade), and the estimate is shown in the column headed "High." An *na* in the "Low" column indicates cases of this kind.

4.3.3 Reciprocal Determinism

The parameters bearing on the issue of reciprocal determinism are those in the columns headed "MEff" and "MAch." Note that the absolute size of these metric estimates varies considerably between the two equations as a function of differences in the scales of self-efficacy and achievement. Self-efficacy scores are standardized internationally to a mean of zero and standard deviation of 1 (OECD, 2005b, p. 369). By contrast, the achievement measure is standardized internationally to a mean of 500 with a standard deviation of 100 (OECD, 2005a, p.131). The parameter estimates shown indicate that the hypothesized reciprocal determinism of mathematics self-efficacy and achievement is supported in 24 of these 30 nations. In each of these cases, the effect of mathematics self-efficacy on achievement and the effect of mathematics achievement on self-efficacy both reach statistical significance. Where the estimates did not support reciprocal determinism, no consistent interpretation was apparent across the countries. In Indonesia and Thailand neither effect

Table 4.2 Estimates of metric coefficients for both structural equations by country

	Achievement equation					Self-efficacy equation		
				Grade				
Country	MEff	SES	Family	Low	High	MAch	Gender	SES
Australia	24.157*	1.428*	17.828*	–41.250*	30.202*	0.006*	–0.335*	0.001
Austria	29.269*	1.471*	7.608*	–72.116*	15.001*	0.004*	–0.424*	0.003
Belgium	52.841*	1.136*	14.934*	*na*	79.703*	0.003*	–0.315*	0.004*
Brazil	91.525*	0.980*	–0.782	–65.657*	24.843*	0.002*	–0.222*	0.003*
Canada	37.406*	0.886*	11.517*	*na*	45.631*	0.006*	–0.301*	0.003*
Czech Republic	39.865*	1.651*	3.071	*na*	28.746*	0.001	–0.362*	0.013*
Finland	16.879*	1.090*	6.949*	*na*	45.280*	0.006*	–0.489*	0.003*
France	56.607*	0.672*	2.885	–34.815*	63.922*	0.006*	–0.254*	0.002*
Germany	43.939*	1.387*	–2.409	–58.452*	36.758*	0.005*	–0.366*	0.000
Greece	56.522*	0.919*	10.778*	–44.916*	11.248*	0.005*	–0.293*	0.004*
Hong Kong-China	31.440*	0.708*	19.275*	–44.942*	27.057*	0.004*	–0.300*	0.004*
Hungary	28.942*	1.798*	7.636*	–56.103*	27.441*	0.007*	–0.306*	0.001
Indonesia	118.422	0.735*	13.556*	–37.365*	42.601*	–0.001	–0.063*	0.002*
Italy	80.603*	0.824*	3.811*	*na*	42.894*	0.005*	–0.203*	–0.001
Latvia	23.405	0.963*	1.833	–51.884*	34.876*	0.005*	–0.244*	0.002
Liechtenstein	47.303*	1.645*	8.568	–42.205*	46.348*	0.006*	–0.445*	–0.004
Luxembourg	45.113*	1.113*	6.479*	–19.480*	46.747*	0.006*	–0.348*	0.000
Macao-China	75.313*	0.318	–1.490	–46.137*	23.732*	0.005*	–0.236*	0.000
Netherlands	19.799*	1.510*	19.579*	*na*	58.092*	0.004*	–0.502*	0.001
New Zealand	31.463*	1.213*	19.010*	*na*	54.609*	0.005*	–0.277*	0.004*
Russian Federation	43.188*	0.734*	1.329	*na*	28.798*	0.004*	–0.237*	0.005*
Slovak Republic	52.322*	1.154*	2.762	*na*	19.017*	–0.003	–0.371*	0.021*
Spain	62.462*	0.510*	2.739	*na*	58.019*	0.005*	–0.207*	0.002*
Switzerland	35.503*	1.167*	13.574*	–63.381*	21.870*	0.005*	–0.449*	0.003*
Thailand	–45.539	1.930*	16.586*	*na*	36.123*	0.002	–0.093*	0.006*
Tunisia	76.426*	0.208	4.316	–26.771*	40.530*	0.005*	–0.196*	0.003*
Turkey	77.768*	0.975*	6.945*	–51.168*	5.431	0.003*	–0.249*	0.010*
United Kingdom	15.819	1.673*	9.919*	*na*	12.923*	0.006*	–0.331*	0.000
United States	43.196*	1.074*	22.888*	–24.496*	10.811*	0.006*	–0.179*	0.003*
Uruguay	70.057*	0.695*	0.592	–34.002*	60.923*	0.004*	–0.239*	0.002*

Note: From Williams and Williams (2010). Copyright 2010 by American Psychological Association

Parameter estimates exceeding 1.96 times their corrected standard error are indicated by an asterisk

MEff mathematics self-efficacy, *SES* socioeconomic status, *Family* family structure, *Low (grade)* lowest of the three grades when students were spread across three grades, *High (grade)* highest grade when students were spread across either two or three grades, *MAch* mathematics achievement

*$p<.05$

was significant. In the Czech Republic and the Slovak Republic, while there was evidence of an effect of self-efficacy on achievement, there appeared to be no significant feedback from achievement. And, in Latvia and the U.K. the reverse situation held; achievement had an effect on self-efficacy but there was no feedback from self-efficacy to achievement. A more detailed interpretation of these coefficients can be found in Williams and Williams (2010).

As one might expect, the estimated effects of self-efficacy on achievement in this situation were different from those obtained in conventional models that ignore the feedback from achievement. As a point of comparison, a conventional model using the same set of variables, and a ‘path analysis’ recursive configuration in which self-efficacy was treated as a variable intervening between the exogenous variables and achievement, was estimated. While direct comparisons of estimates are not legitimate, some observations are possible. First, in the case of the traditional recursive model, the effect of self-efficacy on achievement is significant in each country. Second, comparing this effect estimate with those reported in Table 4.2 above indicates that, where feedback from achievement to self-efficacy is allowed, the traditional one-way effect estimate for self-efficacy on achievement is reduced in 18 of the 30 nations by anywhere from 10% to 70%. In the remaining 12 nations traditional effect estimates are increased by some 10% to more than 100%. In short, one comes to a different view of the influence of self-efficacy on achievement in circumstances where feedback from achievement is provided for.

4.3.4 Other Influences on Mathematics Self-efficacy and Achievement

The effects of socioeconomic status on achievement are positive and statistically significant in all but two nations (Macau-China and Tunisia). Family structure effects on achievement, though modest, reach statistical significance in 18 nations. The effects of the student’s grade-level on achievement are consistent and substantial, pointing to the net advantage of having the opportunity to learn the material assessed. Other things equal, differences between grade-levels amount to some 40 points on average. Gender differences in mathematics self-efficacy are statistically significant and negative in each country, supporting prevailing notions of the gender stereotyping of mathematics performance.

4.4 Modeling Reciprocal Determinism in PISA

The primary purpose of these analyses was to demonstrate a methodology that would allow the statistical modeling of reciprocal determinism with cross-sectional data. The matter of identifying theoretically appropriate variables to be used as instruments was seen as a central issue. The actual model estimated was fairly simple in its structure and restricted by the limited number of degrees of freedom at hand. However, the analyses draw strength from the fact that the model fits the data in 90% of the countries, and supports the proposition of reciprocal determinism in 80%.

(An explanation for the lack of fit in three countries, and the departure from reciprocal determinism in six countries, is beyond the scope of the present analyses but something worth exploring in its own right.)

And, the findings were not without substantive significance. First, overall the analyses are consistent with Bandura's contention that that self-beliefs and performance iteratively modify each other until the individual comes to a realistic appraisal of their self-worth and/or competence relative to the (mathematics) tasks at hand (Bandura, 1978). Second, the replication of this finding in 24 nations suggests that the reciprocal determinism of mathematics self-efficacy and performance may well be a fundamental psychological process that transcends most national/cultural boundaries (Bandura, 2002). Third, the importance of opportunity to learn receives consistent support in these analyses. Other things equal, one year's schooling is worth about one-half of a standard deviation in achievement (45 points on average). Fourth, the consistent negative effect of gender on self-efficacy offers something similar in the way of cross-national consistency in gender stereotyping. Girls everywhere, it seems, underrate their capabilities in mathematics relative to boys with the same degree of mastery.

That said, it is timely to consider some of the methodological limitations associated with nonrecursive models. First, it is important to keep in mind that these analyses do not capture the dynamics of reciprocal determinism. They represent a snapshot of the system at one point in time. Causal lag is not explicitly represented in the model but is assumed to exist. This, in turn, requires the assumption that the system has reached equilibrium. When the system is in equilibrium, the variances and covariances of the variables in question, and the structural and measurement attributes of the model, are unchanging (Kessler & Greenberg, 1981, p.103) and the covariances of variables at time2 add no new information to that available from time1. Given this, the effects between self-efficacy and performance estimated at a single time point are the analogues of cross-lagged reciprocal effects. Since statistical indexes of system stability (Bentler & Freeman, 1983) seem to have limited utility (Kaplan, Harik, & Hotchkiss, 2001, p.338), system equilibrium remains a fundamental but untested assumption for cross-sectional models. In most cases this assumption will have to depend on substantive arguments alone.

In this context it is worth keeping in mind that cross-lagged models have problems of their own. The apparent value of being able to incorporate an explicit time interval in the model may itself be problematic in that it is assumed to approximate the causal interval. To the extent that this lag is misspecified, "parameter estimates may have the wrong sign and be badly biased in magnitude" (Kessler & Greenberg, 1981, p.99). Since social science theory rarely makes concrete predictions about the duration of causal effects, this may be descriptive of most situations. Further, the 'causal interval' in panel studies is often influenced by administrative considerations (the beginning and end of the school year, for example) that have nothing to do with the causal interval.

Kessler and Greenberg (1981, p.28) highlight the essential operational distinction between cross-lagged and cross-sectional models: "The advantage of panel analysis over cross-sectional analysis, then, will not lie in our being relieved of the necessity to make some causal assumptions, but in the possibility of making weaker assumptions than are required with cross-sectional data." They go further and offer the following advice: "the possibility of using cross-sectional or trend data should be considered carefully. If a causal model can be plausibly identified with cross-sectional data, then a panel study may be unnecessary." (Kessler & Greenberg, 1981, p.175).

In order to exploit the unique cross-national comparisons available through the PISA data with respect to the effects of student learning strategies on achievement, and perhaps with respect to the reciprocal determinism of other influences on achievement, nonrecursive models of the kind discussed above seem to have potential. They have their difficulties, mainly in the form of identifying appropriate instrumental variables, and their limitations in the form of system equilibrium assumptions, but they do offer a statistical model that allows for the explicit incorporation of the notion of feedback into considerations about the way in which students learn.

References

Artelt, C., Baumert, J., Julius-McElvany, N., & Peschar, J. (2003). *Learners for life: Student approaches to learning; Results from PISA 2000*. Paris: OECD.

Bandura, A. (1978). The self system in reciprocal determinism. *American Psychologist, 33*, 344–358.

Bandura, A. (2002). Social cognitive theory in cultural context. *Applied Psychology: An International Review, 51*, 269–290.

Bentler, P. M., & Freeman, E. H. (1983). Tests for stability in linear structural equation systems. *Psychometrika, 48*, 143–145.

Berry, W. (1984). *Nonrecursive causal models*. Beverly Hills, CA: Sage Publications.

Bielby, W.T., & Hauser, R.M. (1977). Structural equation models. In A. Inkeles (Ed.), *Annual Review of Sociology, 3* (pp.137-161). Palo Alto, CA: Annual Reviews, Inc.

Carroll, J. (1963). A model of school learning. *Teachers College Record, 64*, 723–733.

Duncan, O. (1975). *Introduction to structural equation models*. New York: Academic.

Fisher, F. (1971). The choice of instrumental variables in the estimation of economy-wide econometric models. In H. M. Blalock Jr. (Ed.), *Causal models in the social sciences* (pp. 245–272). Chicago: Aldine-Atherton.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances*. Baltimore, MD: The Johns Hopkins Press.

Heise, D. (1975). *Causal analysis*. New York: Wiley.

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., & Hopp, C. (2006). Gender comparisons of mathematics attitudes and affect. *Psychology of Women Quarterly, 14*, 299–324.

Kaplan, D. (2000). *Structural equation modeling*. Thousand Oaks, CA: Sage.

Kaplan, D., Harik, P., & Hotchkiss, L. (2001). Cross-sectional estimation of dynamic structural equation models in disequilibrium. In R. Cudek, S. Du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 315–340). Lincolnwood, IL: Scientific Software International, Inc.

Kessler, R., & Greenberg, D. (1981). *Linear panel analysis*. New York: Academic.

Kirsch, I., de Jong, J., LaFontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change: Performance and engagement across countries*. Paris: OECD.

Kline, R. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: The Guilford Press.

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective. *Perspectives on Psychological Science, 1*(2), 133–163.

Martens, M. P., & Haase, R. F. (2006). Advanced applications of structural equation modeling in counseling psychology research. *The Counseling Psychologist, 34*, 878–911.

Maruyama, G. (1998). *Basics of structural equation modeling*. Thousand Oaks, CA: Sage Publications.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton, FL: CRC Press.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide*, 6th edn. Los Angeles, CA: Muthén & Muthén.

Organization for Economic Cooperation and Development. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris: OECD.

Organization for Economic Cooperation and Development. (2005a). *PISA 2003 technical report*. Paris: OECD.

Organization for Economic Cooperation and Development. (2005b). *PISA 2003 data analysis manual*. Paris: OECD.

Organization for Economic Cooperation and Development. (2007). *PISA 2006: Science competencies for tomorrow's world*. Paris: OECD.

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*, 667–686.

Pintrich, P. R., & Maehr, M. L. (2002). *Advances in motivation and achievement: New directions in measures and methods*. Oxford: Elsevier Science.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research and applications* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Stapleton, L. M. (2008). Variance estimation using replication methods in structural equation modeling with complex sample data. *Structural Equation Modeling, 15*, 183–210.

Williams, T., & Williams, K. (2010). Self-efficacy and performance in mathematics: Reciprocal determinism in 33 nations. *Journal of Educational Psychology, 102*, 453–466.

Wright, S. (1960). The treatment of reciprocal interaction, with or without lag, in path analysis. *Biometrics, 16*, 423–445.

Chapter 5
The Measurement of Translation Error in PISA-2006 Items: An Application of the Theory of Test Translation Error

Guillermo Solano-Flores, Luis Ángel Contreras-Niño, and Eduardo Backhoff

Abstract We examined the translation of PISA test items based on the theory of test translation error (TTTE), which has proven to allow detection of translation errors with unprecedented levels of detail. Translation error (TE) is defined as the lack of equivalence between the original and translated versions of items on multiple translation error dimensions (TEDs) that involve design, language, and content. According to the theory, TE results not only from poor translation, but also from factors that are beyond the translators' skills (e.g., languages encode meaning in different ways). We examined the Mexican, Spanish language translation of science and mathematics PISA 2006 items. A panel comprising teachers, translators, a linguist, a test developer, and a measurement specialist examined the translation of 193 text analytical units (55 pieces of introductory text and 138 items) and identified and coded the TEs identified on ten TEDs. For each item, TE was measured as the number of different TEDs on which the review panel identified TEs. To determine which TEDs are critical to student performance, we examined the correlation between TE and item difficulty (percentage of correct answers and mean proportional score, respectively for dichotomous and non-dichotomous items) considering different sets of TEDs. The highest correlations were observed for the sets that included the dimensions, Grammar, Semantics, Register, Information, Construct, and Culture. We also observed different magnitudes of correlations for science and mathematics items and a stronger, statistically significant correlation for translated items whose translation the review panel identified more objectionable than for the rest of the items.

G. Solano-Flores, Ph.D. (✉)
School of Education, University of Colorado at Boulder,
249 UCB, Boulder, CO 80309, USA
e-mail: Guillermo.Solano@Colorado.Edu

L.Á. Contreras-Niño, Ph.D. • E. Backhoff, Ph.D.
Instituto de Investigación y Desarrollo Educativo, University of Baja California, Mexico,
Km. 103 Carretera Tijuana-Ensenada, c.p. 22830 Ensenada, BC, Mexico
e-mail: angel@uabc.edu.mx; backhoff@uabc.edu.mx

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_5,

These results confirm that language- and content-related TEs may threat the validity of translated items. They speak to the value using the TTTE as a formative evaluation tool that PISA countries can use to operationalize translation guidelines.

Keywords Test translation • Test review • Theory of test translation error

Increased awareness of the tremendous sensitivity of tests to language (e.g., Allalouf, 2003; Ercikan, 1998; Ercikan, Gierl, McCreith, Puham, & Koh, 2004; Gierl, Rogers, & Klingner, 1999; Grisay, 2007) in the context of international test comparisons has resulted in recent years in substantial improvements of test translation and adaptation procedures used by PISA (e.g., Grisay, 2003; Harkness, van de Vijver, & Mohler, 2003). As part of these improvements, revised sets of test translation guidelines (e.g., Halleux-Monseur, 2008; Hambleton, 1994; Hambleton, Merenda, & Spielberger, 2005; van de Vijver & Poortinga, 2005) have been made available for participating countries.

Unfortunately, whereas these revised procedures and guidelines are necessary, their implementation and interpretation may not be optimal without procedures that allow countries to perform detailed, systematic evaluations of their own translation work. Available evidence from research on the testing of linguistically diverse populations indicates that, in the absence of tools for systematically examining and discussing the linguistic features of items, reviewers may not be able to detect potential linguistic challenges of those items if they rely solely on their judgment (Solano-Flores & Gustafson, 2012; Solano-Flores, Trumbull, & Kwon, 2003).

This need for conceptual tools in test translation led us to propose a theory of test translation error (TTTE; Solano-Flores, Backhoff, & Contreras-Niño, 2009) which defines translation error (TE) (Note 1) as the lack of equivalence between the original language version and the translated version of items. This lack of equivalence can be examined along multiple dimensions that have to do with the design or visual layout of the items (e.g., format, style), their linguistic features (e.g., grammar, syntax), and their content (e.g., information, construct).

The theory postulates that error in the translation of tests is inevitable. In addition to a poor translation job, TE is due to factors that are beyond the translators' skills. For example, languages encode meaning differently and have different sets of grammatical rules. In addition, TE is multidimensional—an error may involve multiple aspects of language (e.g., the lack of. a comma is a punctuation error but it also may be an error that alters the intended meaning of a sentence). Due to these reasons, and given the linguistic characteristics that are typical among test items (e.g., limited contextualization, high semantic load of terms, compact sentences), it is virtually impossible to preserve exactly the same meaning and linguistic complexity of items across languages.

The notion of test TE as something that cannot be eliminated but can be minimized should be easy to understand by professionals in the educational measurement community. As with measurement error, TE is due to multiple factors (and their interaction), many of which are beyond control. According to the TTTE, effective test translation can minimize, not eliminate, TE. Flawed translated items have many and/or serious TEs; acceptable translated items have few and/or mild TEs.

Conventional translation review procedures focus on determining whether translated items can be accepted (e.g., Grisay, deJong, Gebhardt, Berezner, & Halleux-Monseur, 2007; Mullis, Kelly, & Haley, 1996). They reflect researchers' and evaluators' tendency to emphasize confirming evidence over disconfirming evidence in hypothesis testing (see Church, 1991; Creswell & Miller, 2000). Unlike conventional translation review procedures, a TTTE-based approach focuses on looking for evidence that disconfirms the notion that the translation of test items is correct. We contend that this approach results in more rigorous translation review procedures.

We have used the TTTE to code errors in translated items and develop measures of TE in those items. Moreover, we have been able to link TE and student performance by correlating item difficulty with measures of TE (Backhoff, Contreras-Niño, & Solano-Flores, 2011; Solano-Flores, Backhoff, & Contreras-Niño, 2006). Our findings have shown consistently that translation review based on the TTTE allows detection of TEs with a level of detail not attained with conventional test translation review procedures (Solano-Flores, Contreras-Niño, & Backhoff, 2005; Solano-Flores, Contreras-Niño, & Backhoff, 2006).

In this chapter, we show how detection and measurement of TE can contribute to improved PISA translation procedures. More specifically, we show how coding and measuring TE based on the TTTE allows identification of serious errors in PISA translated items otherwise regarded as acceptable according to conventional translation verification procedures.

Previous empirical evidence showing the sensitivity to TE of review procedures based on the theory comes from reviews of TIMSS items and relatively small samples of released PISA items (Solano-Flores et al., 2005, 2006). In this study, we reviewed a considerably larger sample of items and took into consideration the structure of many of the PISA items—assessment units consisting of one or several paragraphs with contextual information and one or more items related (see Bybee, McCrae, & Laurie, 2009).

5.1 Theoretical Framework

5.1.1 Definition of Translation Error

The theory of test translation error (TTTE) is not only about errors made in test translation, but also about errors in translated tests. According to our theory (Solano-Flores et al., 2009), *test translation* does not refer exclusively to the action of translating items but also to multiple aspects of the entire process through which translated versions of those items are created. *Translation error* does not result exclusively from poor translation job (e.g., inaccuracy of a chosen term, word-by-word translation, use of false cognates); it also results from factors that are beyond the translators' translation skills.

Table 5.1 Acceptability-objectionability of translated items according to the frequency and severity of test translation errors

	Mild errors	Severe errors
Few errors	Acceptable	Questionable
Many errors	Questionable	Objectionable

An example of these factors is the natural, well-known fact that no two languages in the world encode meaning in the same way (see Greenfield, 1997; Nettle & Romaine, 2000). While the translators' job is to ensure that meaning is preserved in their translations, in some cases this is accomplished at the cost of increasing the amount of text. Unlike other forms of text, this is not trivial matter in tests, which students usually need to respond to within certain time limits. Under these circumstances, a substantial increase in the amount of text in an item needed to express the same idea as in its original version may imply more reading time and a potential impact on the time students are left with to make sense of the item.

Another example of aspects beyond the translators' translation skills has to do with the formatting of translated items. Changes in font size and style, and alterations in the proportion of figures included in test items are not due to the translators' actions yet affect the equivalence between the original and the translated versions of an item.

A third example of aspects beyond the translators' translation skills is the extent to which the items reflect the culture of the target language country. While, technically, the translation of an item may not be flawed, the contextual information used in it may not be as familiar to the population tested in the target language as it is to the population tested in the source language.

5.1.2 Inevitability of Translation Error

As a result of the combination of multiple factors like these, strictly speaking, a translation cannot be expected to be perfect. Indeed, our findings from reviews of translated items show that the majority of translated items have TEs—although they are not necessarily fatally flawed (Backhoff et al., 2011; Solano-Flores et al., 2006; Solano-Flores et al., 2009).

5.1.3 Objectionability of Translated Items

To what extent a translated item is objectionable or acceptable depends on the relation between the frequency and severity of TEs. This relationship is represented in Table 5.1. Acceptable translated items have few mild TEs. Questionable translated items have many mild errors or few severe TEs; they may or may not affect student performance depending on the nature of the TEs, the characteristics of the item, and the characteristics of the linguistic group tested. Objectionable translated items have many and severe TEs; they are very likely to alter the intended meaning of the original item and affect student performance.

5.1.4 Translation Error Dimensions

Our theory postulates the existence of test translation error dimensions (TEDs), grouped in three broad categories, Design, Language, and Content. Each TED comprises several types of TE, as shown in Table 5.2. While it parallels the systems of dimensions and types of TEs used in other investigations (e.g., Backhoff et al., 2011; Solano-Flores et al., 2009), the definitions of TEDs shown in Table 5.2 and the types of TE they comprise were respectively adapted and included with the intent to meet the needs of this particular translation review project.

The TEDs, Style, Format, and Conventions, are grouped in the category, Design. These TEDs have to do with the format, editorial features, and visual layout of translated tests. Convention errors are mainly observed in multiple-choice items. TEs belonging to the category, Design tend to be mild and are unlikely to impact student performance (Note 2).

The TEDs, Grammar, Semantics, and Register, are grouped in the category, Language. These TEDs have to do with the structural and functional aspects of the language used in the translation, the preservation of meaning across languages, and the characteristics of the language usage by the target population in social and instructional contexts.

The TEDs, Information, Construct, Culture, and Origin, are grouped in the category, Content. These TEDs have to do with the ways in which information is presented and how examinees are likely to understand and make sense of items. Unlike TEs belonging to the category Design, TEs belonging to the category, Content tend to alter the structural and functional aspects of language or the ways in which examinees make sense of items. Therefore, they tend to be severe and constitute a threat to the validity of a translated item. The TED, Origin addresses the fact that examining the linguistic equivalence of items allows detection of errors not detected throughout the entire process of test development of the item (Solano-Flores, Trumbull, & Nelson-Barber, 2002). Since Origin errors are not exclusive to the translated version of test items, they are included in the list of TEDs only for conceptual purposes, to allow documentation of any anomalies identified during the process of test translation review.

5.1.5 Translation Error Multidimensionality

The theory postulates that test TE is multidimensional. For example, the inappropriate use of commas in *the panda eats, shoots, and leaves* (when the intended meaning is, *the panda eats shoots and leaves*) (Note 3) is both a punctuation error (Style TED) and an error that affects the meaning of the sentence (Semantics TED).

Table 5.2 Translation error dimensions and types of translation errors (italics) considered in the analysis of translated PISA textual analytical units (TAUs)

Design dimensions

Style: The style used in the translation of the TAU is not used in printed materials in the country.

• *Punctuation* • *spelling* • *wrong use of uppercase letter* • *wrong use of lowercase letter*

Format: The visual layout of the translated TAU is different from the original.

• *Change of size, position, or style of an illustration, table, or graph* • *change of justification, font, or font size of text* • *change of margin width* • *omission of graphic component* • *insertion of graphic component*

Conventions: The translation of the TAU does not reflect item writing conventions used in the country.

• *Inconsistent syntactical structure of stem and options* • *wrong use of punctuation in the item's stem* • *change in order of options* • *inconsistent syntactical structure among options* • *wrong use of uppercase letters in options*

Language dimensions

Grammar: The translation of the TAU violates grammatical rules or uses grammatical structures that are not common in the country.

• *Literal translation* • *unnatural syntax of a sentence* • *subject-verb inconsistency* • *singular-plural inconsistency* • *wrong preposition* • *wrong tense* • *conflation of sentences*

Semantics: The translation of the TAU alters its original meaning.

• *Use of a false cognate* • *wrong translation or adaptation of an idiomatic expression* • *alteration of meaning* • *confusing translation of a sentence* • *multiple possible interpretations of a sentence* • *change of gender of a character* • *conflation of ideas* • *inaccurate terms* • *use of terms with multiple meanings* • *wrong translation of a word*

Register: The translation of the TAU does not reflect the terms, idiomatic expressions, and discursive forms used in the country.

• *Use of words of low frequency in the country* • *wrong translation of a technical term* • *translation of a technical term in a way not used in the country*

Content dimensions

Information: The translation of the TAU alters the amount, precision, or type of information provided.

• *Inconsistent translation of a non-technical term* • *change in number of times a technical term is used* • *insertion of technical term* • *insertion of a sentence or explanation* • *omission of a key word* • *omission of a technical term* • *omission of a sentence or explanations*

Construct: The type of skill or knowledge needed to understand and respond to the TAU is different from the skill or knowledge needed to understand and respond to the TAU in the source language.

• *Possible change of the item's cognitive demands* • *possible alteration of ways in which a task may be interpreted* • *wrong technical term* • *inconsistent translation of a technical term* • *undue insertion of a technical term* • *omission of a technical term* • *translation of a technical term as a non-technical term* • *translation of a non-technical term as a technical term*

Culture: The TAU does not reflect the characteristics of the culture or the curriculum in the target language.

• *Contextual information and situations that are uncommon in the country* • *measurement units not used in the country* • *problem posed not meaningful in the country's culture* • *knowledge assessed not taught in country*

Origin: The TAU carries over errors from the source language version.

• *Inconsistency in the content of the two source languages* • *conceptual errors in the design of the item* • *confusing directions* • *the answer to an item may give the clue for responding to another item within the same assessment unit*

5.1.6 Tension Among Translation Error Dimensions

Finally, the theory postulates that there is a tension between TEDs. Actions intended to avoid TE on a given TED may involve making errors on other TEDs. For example, the grammatical rules of the target language may prevent a noun from being repeated in the same sentence. In some languages, a marker needs to be used to refer to a noun in the rest of the sentence, once the noun appears in it. As a consequence, a key technical term that appears several times in the same sentence in the original version of the item appears only once in its translation. The grammatical rules of the target language need to be followed at the cost of altering the number of times that the key term appears in the sentence—which alters the amount of information provided by the item.

5.2 Methods

5.2.1 Sample of Assessment Units and Analytical Test Units

We examined 61 assessment units (one or several paragraphs with contextual information and one or more items related) from the Mexican, Spanish language version of PISA-2006. Of these 61 assessment units, 37 and 24 were respectively science and mathematics assessment units (Note 4). These 61 assessment units comprised a total of 193 text analytical units (TAUs), defined as either the introductory text or an item within an assessment unit. Of the 193 TAUs examined, 55 were introductory texts and 138 were items. Of these 138 items, 101 and 37 were respectively science and mathematics items.

5.2.2 Test Translation Review and Error Coding Procedures

In addition to the fact that most of the PISA 2006 items consisted of two forms of TAUs (an introductory text or an item), our coding procedure took into account that PISA items use two source languages, English and French (see Grisay et al., 2007).

We assembled a multidisciplinary translation review panel composed of three middle school teachers (Spanish, science, and mathematics); three high school teachers (Spanish, science, mathematics); one English-to-Spanish translator, and one French-to-Spanish translator (both certified by international translation professional organizations); one linguist; one test developer; and one psychometrician (measurement specialist).

The following procedure was used to review each TAU. First, the TAU in the target language (the translated item) was projected on a screen. Reviewers read the TAU and, in the case of items, responded to the item individually as if they were students taking the test. This was done with the purpose of giving the reviewers the opportunity to become acquainted with the content of the item and to become aware of its cognitive and linguistic demands in the target language.

The reviewers then were asked to examine the TAU and individually record on a coding form all the types of TE they thought could affect the interpretation of the item. The reviewers were instructed to focus on a specific set of dimensions designated according to their professional background. However, they were allowed to record errors on all dimensions (Table 5.3).

Once the reviewers finished recording their comments, the original English and French versions of the TAU were projected on two additional screens. Then the reviewers were asked to compare the English and French versions with the TAU in the target language and to individually code any type of TE according to the list of types of errors listed above for each error dimension. They also wrote their comments on the TAU based on their experience reading and responding to it and on comparing the original and translated versions.

For each TED, the panel discussed each reviewer's coding. Project staff facilitated a discussion to ensure that the panel decided by consensus what errors should be recorded and on which TEDs they should be coded. In the case of items, the panel was asked to decide, based on the number and severity of the TEs, if the translated item should be classified as objectionable (i.e., an item with many and severe TEs which were likely to adversely affect student performance). The review coding decisions were captured on an electronic spreadsheet for further analysis.

5.2.3 Data Analysis

For the purpose of our analysis, we measured TE in each TAU as the number of different translation error dimensions (NDTED) on which TEs were observed in it. This coarse-grain measure has proven to be sensitive to important differences in translation quality among items (see Solano-Flores et al., 2005, 2006).

Also for the purpose of our analysis, we used the p-values of items as a measure of item difficulty. Item p-value was computed as the proportion of the item's highest possible score (see Adams, Berezner, & Jakubowski, 2010), which allowed to have proportional measures of difficulty for both dichotomous and partial-credit items. More specifically, for dichotomous items, difficulty was computed as the proportion of students who responded correctly; for partial-credit items, difficulty was computed as the mean score of the item divided by its maximum score.

To examine the impact of TE on student performance, we examined the Pearson correlations between NDTED and item p-value for different sets of TEDs, different content areas (science and mathematics), and items that were and were not identified as objectionable by the translation review panel. Impact on performance should be observed as a negative correlation.

Given the complex interaction of the students' knowledge of the content being assessed and the cognitive and linguistic demands of test items, it would be naive to

Table 5.3 Expertise provided and assigned focus on translation error dimensions by specialist: 1 = main role; 0 = adjuvant role

Expertise and Contribution	Style	Format	Conventions	Information	Grammar	Semantics	Construct	Register	Culture	Origin
Teacher: Formal and informal use of language in the school; consistency of TAUs with language used in the curriculum; content accuracy.	0	1	1	1	0	0	0	1	1	0
Translator: Language equivalence across languages; formal use of language in the translated TAUs.	1	1	0	1	1	1	0	0	0	0
Linguist: Formal (structural) and functional (sociolinguistic, cultural) aspects of the translated TAUs; use of language in context.	1	0	0	0	1	1	1	0	1	0
Test developer: Item writing; equivalence of TAUs; item complexity across languages; wording.	0	1	1	1	0	0	1	0	0	0
Psychometrician: Potential impact of language and design features on the interpretation of TAUs.	0	0	1	1	0	1	1	0	0	0

Table 5.4 Percentage of text analytical units (n = 193) with at least one error on each of the translation error dimensions

Dimension	Percent
Style	48
Format	53
Conventions	3
Information	53
Grammar	53
Semantics	78
Construct	35
Register	21
Culture	5
Origin	41

expect to observe impressively high and statistically significant correlations. Rather, we expected to observe patterns in those correlations that would indicate a systematic impact of TE on student performance, especially for language- and content-related TEDs and for items identified as objectionable by the translation review panel.

5.3 Results

5.3.1 Frequency and Severity of Translation Errors

We observed TEs on at least one dimension for almost all (96%) of the TAUs. Of the 138 TAUs which consisted of items, 26 were identified by the committee as objectionable.

Table 5.4 shows the percentage of TAUs identified as having at least one error on each of the TEDs. As indicated above, many of these errors are not likely to bias test results and many are even difficult to be noticed by individuals who have no experience reviewing test translations. On the other hand, there are TEs that may potentially threaten the validity of test items. Such is the case for errors on the Semantics, Grammar, and Information dimensions, which were observed respectively in 78%, 53% and 53% of the TAUs.

On average, a TAU had errors on 3.9 different dimensions (s.d. = 1.834). As Fig. 5.1 shows, the number of different dimensions in which error was observed had a normal frequency distribution.

5.3.2 Translation Error and Item Difficulty

As Table 5.5 shows, Pearson correlation coefficients of −.059 and −.117 between NDTED and p-value were observed respectively when all dimensions were considered and when the three language dimensions (Grammar, Semantics, and Register) and three of the four content dimensions (Information, Construct, and Culture)

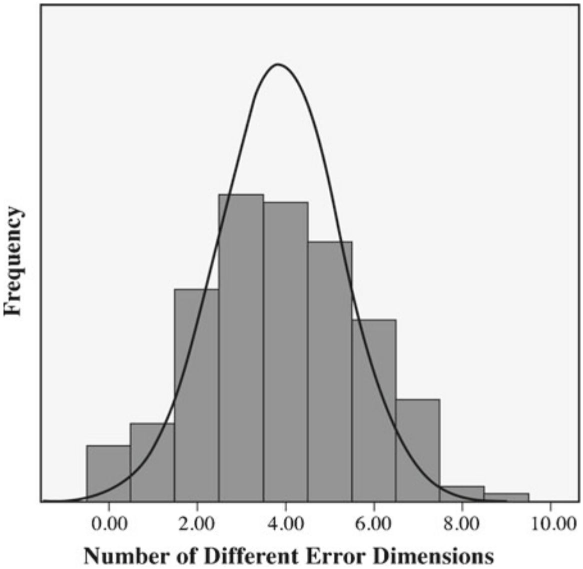


Fig. 5.1 Frequency distribution of text analytical units by number of different error dimensions

Table 5.5 Correlation between number of different dimensions on which error was observed and item p-value by set of dimensions considered, content area, and objectionability

Comparison	Correlation
By set of dimensions (138)	
All dimensions	–.059
Language and content dimensions[a]	–.117
By content area[a] (138)	
Science (n = 101)	–.115
Mathematics (n = 37)	–.213
By objectionability (language and content dimensions)[a](138)	
Non-objectionable items (n = 112)	–.084
Objectionable items (n = 26)	–.404**

Sample and subsample sizes in parentheses

[a] Includes the three language dimensions (Grammar, Semantics, and Register) and three of the four content dimensions (Information, Construct, and Culture)

**Significant at p = .01 (2-tailed)

were considered. (As mentioned above, since Origin errors are common to both the source and language versions of items, they were not included in the analyses). This difference supports findings from previous test translation reviews that design dimensions (Style, Format, and Conventions) are unlikely to affect student performance whereas language and content dimensions tend to have a greater impact on student performance and may potentially threaten the validity of translated items.

Correlation coefficients of −.115 and −.213 between NDTED and p-value were observed respectively for the science and mathematics items. These results are consistent with findings from other translation reviews, in which we (e.g., Solano-Flores, Backhoff, & Contreras-Niño, 2005) have observed higher correlations between NDTED and item difficulty for mathematics than science items.

Correlation coefficients of −.084 and −.404 (significant) were observed respectively for acceptable and objectionable items. This considerable difference indicates that the review procedure allows identification of items which have sets of errors that are likely to seriously impact student performance. This finding is important, considering that the number of items identified as objectionable (26) constitute about 19% of the 138 items examined.

5.4 Summary and Conclusions

The theory of test translation error (TTTE; Solano-Flores et al., 2009) postulates the existence of translation error dimensions (TEDs; e.g., Semantics, Construct, Grammar) and views translation error (TE) as multidimensional (a translation error can belong to several TEDs). It also postulates that a tension exists between TEDs (i.e., in translating a test item, avoiding error on one dimension may produce error on other dimensions). Accordingly, error-free test translation is impossible; effective test translation minimizes but does not eliminate error. The theory also postulates that while items usually have multiple TEs, most of them are mild and even unnoticeable. Objectionable translated items have many and severe TEs and are likely to pose serious linguistic challenges to examinees who are given the translated version of a test.

In this chapter, we report the results of our review of the Spanish language Mexican version of PISA-2006 science and mathematics text analytical units (TAUs). Consistent with results from our review of the Spanish Mexican translation of TIMSS-1995 (Solano-Flores et al., 2005) and the Spanish Mexican translation of PISA-2003 (Backhoff et al., 2011), our results show that translation reviews based on the TTTE are highly sensitive to TE.

The results also confirm previous findings that student performance tends to be resilient to TE on design-related TEDs and sensitive to TE on language- and content-related TEDs. Also, items whose translation was identified as objectionable by the review panel correlated higher with item difficulty than items whose translation was not identified as objectionable—a finding that speaks to the sensitivity of TTTE-based judgmental review procedures.

A limitation of our analyses of correlations of measures of TE and item difficulty stems from the fact that we did not account for the effect of TE observed in the introductory text of assessment units. Future research should explore models for examining this relationship.

Unlike other approaches created to examine translation quality, the TTE focuses on disconfirming (rather than confirming) evidence that the translation of test items is correct. In addition, because they use multidisciplinary review panels which discuss the linguistic features of the items at length, TTTE-based coding procedures are sensitive to TE with a level of precision and detail not attained with conventional approaches.

Experienced test translators who have attended our workshops on the use of the TTTE and the methods described in this chapter (e.g., Backhoff, Solano-Flores, & Contreras-Niño, 2010; Solano-Flores et al., 2010) react initially with skepticism when we report our findings. They find it difficult to believe that items translated according to available translation guidelines have multiple TEs. It is not until they observe the discussions of the review panels examining specific translated items that they appreciate the level of sensitivity of the theory and our coding procedures to the nuances of language in translated items.

As with measurement error, TE cannot be entirely eliminated, but it can be minimized. As our results show, a theoretical perspective that assumes error inevitability in test translation is more sensitive to the complexities of language in translated PISA items and can contribute to the improvement of future PISA translation procedures. We hope that, in the future, PISA participating countries use our approach as a tool for operationalizing PISA translation procedures and formatively evaluating their own translation work.

Author's Note

Portions of this paper are originally from a paper presented at the PISA Research Conference, 14–16 September 2009, Kiel, Germany. The investigation reported in this paper was commissioned and funded by the National Institute for Educational Evaluation (INEE), Mexico, and conducted through a contract with the Autonomous University of Baja California (UABC), Mexico. The opinions expressed are not necessarily those of the funding agency. Contact author: Guillermo Solano-Flores, guillermo.solano@colorado.edu.

Notes

Note 1. While *translation error* (in singular) is used here to refer to lack of equivalence between the original language version and the translated version of an item, *translation errors* (in plural) or *a translation error* are used to refer to specific instances or types of translation error (e.g., the inaccurate translation of a term or an inappropriate use of punctuation).

Note 2. Of course, there are exceptions. For example, an alteration in the proportion of the length of the axes in a graph showing a functional relationship may make the line of the function look steeper in the translated item than in the original—which may affect how the examinee interprets the function.

Note 3. The example is based the story told by Lynne Truss (2004) at the beginning of her well-known book on punctuation, *Eats, shoots, and leaves*.

Note 4. One of the science assessment units and 17 of the mathematics assessment units consisted of a stand-alone item with no introductory text.

References

Adams, R., Berezner, A., & Jakubowski, M. (2010). *Analysis of PISA 2006 preferred items ranking using the percent-correct method* (OECD Education Working Papers, No. 46). OECD Publishing. Retrieved June 7, 2011, http://dx.doi.org/10.1787/5km4psmntkq5-en

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education, 16*(1), 55–73.

Backhoff, E., Contreras-Niño, L. A., & Solano-Flores, G. (2011). *The theory of test translation error and the TIMSS and PISA international test comparisons*. Mexico: National Institute for Educational Evaluation [Sp.].

Backhoff, E., Solano-Flores, G., & Contreras-Niño, L. A. (2010, February 18–19). *Analysis of the Mexican Spanish language translation of PISA-2006*. Presentation at the Ibero-American Seminar on the theory of test translation error in international comparisons. National Ministry of Education and National Institute for Educational Evaluation, Mexico City, Mexico.

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching, 46*(8), 865–883.

Church, B. (1991). An examination of the effect that commitment to a hypothesis has on auditors' evaluations of confirming and disconfirming evidence. *Contemporary Accounting Research, 7*(2), 513–534.

Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into Practice, 39*(3), 124–130.

Ercikan, K. (1998). Translation effects in international assessment. *International Journal of Educational Research, 29*, 543–553.

Ercikan, K., Gierl, M. J., McCreith, T., Puham, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301–321.

Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Using statistical and judgmental reviews to identify and interpret translation DIF*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, QC.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52*(10), 1115–1124.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225–240.

Grisay, A. (2007). *The challenge of adapting PISA materials into non Indo-European languages: Some evidence from a brief exploration of language issues in Chinese and Arabic*. OECD Core A Consortium.

Grisay, A., de Jong, J. H., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement, 8*(3), 249–266.

Halleux-Monseur, B. (2008). *Translation, adaptation and verification of test material in OECD international surveys*. Paris: Directorate for Education, Institutional Management in Higher Education Governing Board, OECD.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*(3), 229–244.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Harkness, J., van de Vijver, F. J. R., & Mohler, P. (Eds.). (2003). *Cross-cultural survey methods*. Hoboken, NJ: Wiley.

Mullis, I. V. S., Kelly, D. L., & Haley, K. (1996). Translation Verification Procedures. In M. O. Martin & I. V. S. Mullis (Eds.), *Third international mathematics and science study: Quality assurance in data collection*. Chestnut Hill, MA: Boston College.

Nettle, D., & Romaine, S. (2000). *Vanishing voice: The extinction of the world's languages*. New York: Oxford University Press.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. A. (2006). *Methodology for evaluating the quality of test translations in international test comparisons: The case of Mexico, TIMSS-1995. [Sp.]*. Mexico: National Institute for Educational Evaluation (INEE).

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. A. (2009). Theory of test translation error. *International Journal of Testing, 9*, 78–91.

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. A. (2010, February 18–19). *Test translation review sessions: A demonstration*. Presentation at the Ibero-American Seminar on the theory of test translation error in international comparisons. National Ministry of Education and National Institute for Educational Evaluation, Mexico City, Mexico.

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2005, April 12–14). *The Mexican translation of TIMSS-95: Test translation lessons from a post-mortem study*. Paper presented at the 2005 annual meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2006). Test translation and adaptation: Lessons learned and recommendations for countries participating in TIMSS, PISA, and other international comparisons. *REDIE: Electronic Journal of Educational Research, 8*(2). [Sp.] http://redie.uabc.mx/vol8no2/contents-solano2.html

Solano-Flores, G., & Gustafson, M. (2012). Assessment of English language learners: A critical, probabilistic, systemic view. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large scale assessment in education: Theory, issues, and practice*. New York, NY: Taylor & Francis, Routledge.

Solano-Flores, G., Trumbull, E., & Kwon, M. (2003, April 21–25). *The metrics of linguistic complexity and the metrics of student performance in the testing of English language learners*. Symposium paper presented at the 2003 Annual Meeting of the American Evaluation Research Association. Chicago.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2*(2), 107–129.

Truss, L. (2004). *Eats, shoots & leaves*. New York: Gotham Books.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Chapter 6
PISA Student Nonresponse Adjustment Procedures

Keith Rust, Sheila Krawchuk, and Christian Monseur

Abstract Large-scale surveys experience nonresponse that might result in biases in the results. In educational surveys, nonresponse can occur at three levels, i.e. the school, the student and the item. Traditionally, large-scale surveys in education have compensated for unit nonresponse, both at the school and at the student level, by applying survey weight adjustments.

IEA TIMSS and PIRLS surveys, for example, apply a school nonresponse adjustment calculated within the explicit stratum level and a student nonresponse adjustment calculated within each school. Since its first data collection, PISA has implemented the school nonresponse adjustment at a lower level than the explicit stratum, using combinations of the implicit stratification variables. The student nonresponse adjustment procedures have continuously evolved since the first data collection to better reflect differential response rates. In PISA 2006, the student nonresponse adjustment was computed by creating, per school nonresponse adjustment cell, four student nonresponse adjustment cells based on relative grade level, and gender.

K. Rust, Ph.D (✉)
Westat, 1600 Research Boulevard RE400, Rockville, MD 20850-3129, USA
e-mail: KeithRust@Westat.com

S. Krawchuk, MS
Westat, 3100 Wyntree Drive, Norcross, GA 30071, USA
e-mail: SheilaKrawchuk@Westat.com

C. Monseur, Ph.D.
Département Education et Formation, Université de Liège,
bld. du Rectorat 5 (B32), 4000 Liège, Belgium
e-mail: cmonseur@ulg.ac.be

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_6,

This study compares the efficiency of the 2003 and the 2006 procedures for reducing potential bias due to differential student response rates. The comparison highlights the potential of the 2006 method for reducing bias due to differential response rates according to gender and grade.

Keywords Weighting • Nonresponse bias • Differential response rate

6.1 Nonresponse Adjustments in PISA

Survey weights are required for the analysis of data from a complex sample design, such as PISA. The weights are needed to ensure that bias in estimation is not introduced through differential sampling rates for different parts of the population, whether designed or a result of inaccurate data on the sampling frame. The weights are also adapted so as to minimize, to the extent possible, the biasing effects of nonresponse. For a discussion of the use of weights in survey data analysis, see Kish (1992) and Pfeffermann (1996).

The survey weights attached to student records in the PISA data base, and used in the analyses of PISA data, consist of several components:

1. The school base weight, which is calculated as the reciprocal of the sample selection probability of the school.
2. The student-within-school weight, calculated as the reciprocal of the (conditional) within-school sample selection probability.
3. Trimming factors, potentially at both the school and student level. These are intended to reduce sampling variance that may result from sample selection probabilities which, due to faulty information at the time of sample selection, resulted in unusually large school base weights or student-within-school weights.
4. Adjustments to the school base weight, intended to reduce any bias introduced by school nonresponse.
5. Adjustments to the student base weight (the product of the school base weight, the school trimming factor, the school nonresponse adjustment, and the student-within-school weight), aimed at reducing any bias introduced as a result of student nonresponse.

While the calculation and application of school base weights and student-within-school weights are standard procedures in sample survey applications, the process of deriving suitable nonresponse adjustments is less standard. The relative success of any given approach tends to be very dependent on the pattern of survey nonresponse, and the relationship of that pattern to the key survey data. For discussions of general approaches to deriving survey nonresponse adjustments, see Kalton (1983), Little (1986), Brick and Kalton (1996), and Kalton and Flores-Cervantes (2003). The calculation of trimming factors is also not routine, but we do not discuss that further here. The trimming factors play only a minor role in determining the final student weights in PISA.

The derivation of school nonresponse adjustments in PISA tends to be very situation specific, varying from country to country, and even year to year for a given country. There are two reasons for this. First, school response rates vary considerably across countries and over time. Many countries achieve 100% school response or very close to it; others struggle to reach 70%. Second, the variables that are available for constructing nonresponse adjustments (which must be available for nonrespondents as well as respondents) vary greatly across countries. Even seemingly common variables (public versus private schools, for example) can vary greatly as to their relationship with PISA achievement from country to country.

On the other hand, student nonresponse adjustment procedures tend to be standardized more closely across countries. There are several reasons for this.

1. Student response rates are much more uniform across countries and over time than school response rates. In almost all cases national student response rates for PISA are between 85% and 95%.
2. There are only a limited set of variables available for use in nonresponse adjustments. The PISA data collection procedures lead to the following variables being available for students who do not respond, but no others: school; gender; month of birth; and grade.
3. Much more so than in the case of school nonresponse, the same characteristics tend to be associated with student achievement across countries. Gender and school grade are often strongly related to achievement, particularly grade, which is far from surprising. After controlling for grade however, month of birth (within the 12 months span that defines the PISA population) bears little relationship to achievement.

Thus it would seem reasonable to adopt a standardized approach to the creation of student nonresponse adjustments, based on the student's school, grade, and gender. This paper discusses how PISA has used these variables over its different cycles. In particular, in 2006 the procedure changed from that in 2003, by adding gender, increasing the role of grade, and decreasing the role of school, in deriving student nonresponse adjustments. We evaluate the effect of that change, by applying both methods to the data collected in 2006, and comparing the results. The results were compared both with respect to the distributions of the sample weights by gender and grade for each country considered, and also with respect to the mean national achievement score for mathematics, reading, and science.

A major purpose of the analysis was to determine whether the change in procedures had induced an artifactual trend component to the comparison of 2006 results with those from 2003. Since there are many countries that participated in both these years, we decided to restrict the analysis to this set of countries.

Details of all aspects of the survey weighting procedures used in PISA are contained in the technical reports for the respective cycles (Organisation for Economic Co-operation and Development [OECD], 2002; OECD 2005, 2009).

6.1.1 Student Nonresponse Adjustments in PISA

One of the major technical considerations in deriving student nonresponse adjustments is the determination of what characteristics should be used to define classes of students who will be treated as a single unit for student nonresponse adjustment, so that the students within that class all have their weights increased, by the same ratio, to account for unassessed students in the class. The other major consideration, which is related to this, is to determine the minimum size of such classes. The key considerations in determining how many and which variables to use in forming classes are that (1) the variable must be known for all students, whether they were assessed or not; (2) it should be related to the response rate; (3) it should be related to achievement.

In PISA 2000, generally speaking each school formed such a class. That is, within each school the participating students had their weights increased by a constant factor, to account for nonrespondents within the same school. Clearly school membership generally satisfies the three conditions above (although in a few countries condition 3 holds only weakly, and in others condition 2 may only hold weakly). The other attraction of using school to define these classes is that most schools have a good sample size of students to form the basis of a stable class, and school is what has been used in IEA studies such at TIMSS.

However, following the results of PISA 2000, two other variables, available for all students, were clearly related to achievement – grade level (which could have been anticipated, and is not relevant in IEA studies, which are of a single grade) and gender. Grade level has a strong relationship to both achievement and response in most countries. Thus in 2003 we used high/low grade categories as well as school to determine the student nonresponse classes. This was very effective in a few countries – those where many schools in the sample had a significant number of PISA students from more than one grade, but in fact there were few such countries. This approach was also a little problematic in that typically a given school has a sample not much larger than a reasonable desirable minimum sample size for student nonresponse classes (generally agreed in the survey field to be in the vicinity of 10–20 students), so that adding grade in addition to school tended to create classes that were too small, and thus were subsequently collapsed back to the school level.

Thus the 2003 procedure was not particularly effective at improving on the 2000 procedure. In planning the 2006 weighting process, the PISA consortium realized that if grade and gender are to be meaningfully included in the process of forming weight classes, school per se cannot be used to form classes. Thus the procedure for 2006 was modified to form student nonresponse classes on the basis of school nonresponse class (a group of six or more similar schools from the same explicit stratum), high/low grade categories and gender where possible, and either grade or gender where it is not possible to use both effectively.

The aim was to capture most of the school differences in achievement and student response rate through using the school nonresponse class, rather than the individual school, thus leaving scope for the effective use of at least one of grade and gender, and, where appropriate and feasible, both.

The PISA consortium is confident that this revised procedure led to reduced nonresponse bias in the PISA results. However, for countries that participated in 2003, this raises the issue of whether the change in student nonresponse adjustment procedures could lead to an artifactual impact on trend estimates. This could result because the 2006 estimates are subject to less student nonresponse bias than those of 2003 (and 2000). In this paper we evaluate the extent to which the change in student nonresponse adjustment procedures led to results that differ from those that would have been obtained had the 2003 procedures been retained.

The student nonresponse adjustment was computed in PISA 2006 as follows:

1. Per school, four student nonresponse adjustment cells were created:
   (a) Higher grades/girls;
   (b) Higher grades/boys;
   (c) Lower grades/girls;
   (d) Lower grades/boys,where the high/low grade split was derived within each explicit stratum, and thus varied across countries and even within countries across strata.
2. In single sex schools or in schools with only students attending one grade, only two student nonresponse cells were created.

The two major changes between the previous procedures (used in 2003) and the PISA 2006 procedure were (a) the addition of the gender variable for creating the nonresponse cell, and (b) the ordering of the collapsing. In PISA 2003, nonresponse cells were firstly collapsed within school, and then, if required, schools were collapsed. In 2006, a nonresponse cell from a school was firstly collapsed with a nonresponse cell sharing the same gender and grade but from another school. However, these two schools had to be in the same school nonresponse cell and explicit stratum. If, after collapsing to the level of the school nonresponse adjustment cell, further collapsing was required, usually nonresponse cells were collapsed across gender and then (if necessary) across grade.

As this modification in the computation of the student nonresponse adjustment might have an impact on population estimates, and in particular on performance estimates, it was decided to compute, for the 2006 data, the student nonresponse adjustment according to (i) the PISA 2006 algorithm and (ii) the PISA 2003 algorithm, for the countries that participated in the 2003 and 2006 surveys. Comparing population estimates for the two sets of weights permits measuring the impact of the weighting modification. This in turn permits an evaluation of whether trends from 2003 to 2006 likely reflect real changes within a country, or are an artifact of the enhancement to the weighting procedure.

6.1.2 Comparison of Student Nonresponse Adjustment Methods

Using the data from PISA 2006, we undertook a comparison of the effects of the nonresponse adjustment method that was actually applied in producing the weights for 2006, with the method that had been used in 2003. We compared these as to their effects on the distribution by gender and by grade, and their effects on the mean achievement score for each of mathematics, reading, and science.

Three sets of weights were used in these analyses:

1. the initial student weight that consists of the product of:
   (a) the school base weight;
   (b) the school trimming factor;
   (c) the school nonresponse adjustment factor;
   (d) the student-within-school weight;
2. the final student weight based on the 2003 nonresponse adjustment method;
3. the final student weight based on the 2006 nonresponse adjustment method.

For the second and third sets of weights, only responding students were included in the analyses, while responding and nonresponding students were included for the first set of weights.

Table 6.1 presents the comparison between these different sets of weights on the gender variable. Differences in response rates for boys and girls of more than two percentage points are highlighted, while those differences of between 1.5 and 2.0 percentage points are also shaded.

In seven countries, the difference between the boys' response rate and the girls' response rate is greater than two percentage points, i.e. Austria, Denmark, Iceland, Italy, Poland, Spain, and Tunisia; in each case the response rate for girls was higher than for boys. In these seven countries, the 2006-method weighted estimates are equal or very close to the estimates computed using the initial student weights and data from the Student Tracking Form, while the 2003-method weighted estimates differ to a greater extent. The 2006 adjustment method appears more efficient in reducing a potential bias due to the differential participation rates between boys and girls. However, it is noteworthy that the differences in results between the two methods vary by country. This is a reflection of the fact that in some countries the nonresponse differential by gender is associated with differences by school and grade, so that the 2003 adjustment method is still effective at reducing the gender bias, whereas in other countries it is not. For example, in the case of Austria, there is a large difference in the response rates for boys and girls (4.8 percentage points). Yet the results of the 2003 method give close to the same percentage of students in each gender category as the initial weights (49.3% girls compared to 49.1%). In Iceland, on the other hand, where the difference in response rates between boys and girls is a much smaller 2.6 percentage points, the difference in the percentage of girls between the weights adjusted via the 2003 method, and the initial weights, is larger (50.5% compared to 49.6%).

Table 6.2 presents the results for the grade variable. Only grades that have at least 100 students in sample are reported. Only countries where there is a difference of at least 0.5% between methods, for at least one grade, are presented. Cases are highlighted where either the 2003 or 2006 method results in a difference from the results using the initial weights by one percentage point or more.

Table 6.1 Response rates[a] and weighted percentages by gender, based on (i) initial student weights, (ii) 2003-method adjusted student weights and (iii) 2006-method adjusted student weights

		Percentage girls				Percentage boys			
Country	Response rate – girls	Initial weights	2003 adjustment method	2006 adjustment method	Response rate – boys	Initial weights	2003 adjustment method	2006 adjustment method	Difference in response rate
Australia	85.2	48.8	48.5	48.9	86.4	51.2	51.5	51.1	–1.2
Austria	92.9	49.1	49.3	49.1	88.2	50.9	50.7	50.9	**4.8**
Belgium	93.1	47.6	47.4	47.6	92.9	52.4	52.6	52.4	0.2
Brazil	90.6	53.9	54.0	53.8	90.7	46.1	46.0	46.2	–0.2
Canada	77.9	49.6	49.9	49.7	77.2	50.4	50.1	50.3	0.8
Czech Republic	89.9	43.5	43.3	43.4	90.9	56.5	56.7	56.6	–1.0
Denmark	90.7	50.3	51.2	50.3	87.5	49.7	48.8	49.7	**3.2**
Finland	93.4	50.3	50.6	50.4	92.1	49.7	49.4	49.6	1.3
France	89.3	51.3	51.6	51.5	89.0	48.7	48.4	48.5	0.3
Germany	92.3	48.4	48.7	48.4	91.4	51.6	51.3	51.6	0.9
Greece	95.0	49.8	49.6	49.7	95.2	50.2	50.4	50.3	–0.2
Hong Kong - China	91.6	50.7	50.7	50.7	91.1	49.3	49.3	49.3	0.5
Hungary	93.4	47.9	47.9	47.9	92.8	52.1	52.1	52.1	0.6
Iceland	84.4	49.6	50.5	49.6	81.8	50.4	49.5	50.4	**2.6**
Indonesia	98.8	48.7	49.2	48.7	96.9	51.3	50.8	51.3	1.9
Ireland	83.8	50.6	50.4	50.6	83.7	49.4	49.6	49.4	0.0
Italy	93.2	50.4	50.8	50.4	90.9	49.6	49.2	49.6	**2.2**
Japan	99.5	49.9	49.9	49.9	99.6	50.1	50.1	50.1	–0.1
Korea	99.4	49.3	49.4	49.3	98.7	50.7	50.6	50.7	0.7
Latvia	97.3	51.4	51.7	51.4	96.0	48.6	48.3	48.6	1.2
Liechtenstein	96.8	53.8	54.2	53.8	95.1	46.2	45.8	46.2	1.8
Luxembourg	96.7	49.4	49.3	49.4	96.3	50.6	50.7	50.6	0.4
Macao	98.0	49.4	49.6	49.4	97.1	50.6	50.4	50.6	0.9

(continued)

Table 6.1 (continued)

Country	Response rate – girls	Percentage girls			Response rate – boys	Percentage boys			Difference in response rate
		Initial weights	2003 adjustment method	2006 adjustment method		Initial weights	2003 adjustment method	2006 adjustment method	
Mexico	95.3	52.0	51.9	51.9	95.5	48.0	48.1	48.1	–0.1
Netherlands	90.1	49.1	49.0	49.1	89.7	50.9	51.0	50.9	0.3
New Zealand	87.6	51.6	51.8	51.6	86.4	48.4	48.2	48.4	1.2
Norway	88.8	48.3	48.9	48.3	86.9	51.7	51.1	51.7	1.9
Poland	92.8	50.4	50.9	50.3	90.6	49.6	49.1	49.7	**2.1**
Portugal	87.2	51.5	52.0	51.7	85.6	48.5	48.0	48.3	1.7
RussiaFederation	96.5	51.9	52.2	52.1	95.5	48.1	47.8	47.9	1.1
Serbia	94.1	49.2	49.0	49.2	93.7	50.8	51.0	50.8	0.5
Slovak Republic	93.2	48.6	48.8	48.6	92.9	51.4	51.2	51.4	0.3
Spain	90.3	49.4	50.1	49.4	86.7	50.6	49.9	50.6	**3.7**
Sweden	91.2	48.8	48.7	48.7	91.5	51.2	51.3	51.3	–0.3
Switzerland	95.4	48.4	48.5	48.4	94.6	51.6	51.5	51.6	0.8
Thailand	99.0	57.4	57.7	57.4	98.3	42.6	42.3	42.6	0.7
Tunisia	95.5	52.3	52.8	52.2	93.5	47.7	47.2	47.8	**2.1**
Turkey	98.3	45.3	45.5	45.3	97.0	54.7	54.5	54.7	1.3
United Kingdom	87.0	50.4	50.4	50.5	86.7	49.6	49.6	49.5	0.3
United States	90.9	49.4	49.4	49.4	91.0	50.6	50.6	50.6	0.0
Uruguay	87.2	51.2	51.1	51.2	86.0	48.8	48.9	48.8	1.1

[a] These responses rates were computed using the student initial weight, which includes the school nonresponse adjustment factor

Table 6.2 Response rates and weighted percentages by grade, based on (i) initial student weights, (ii) 2003 method adjusted student weights and (iii) 2006 method adjusted weights

Country	Grade	Response rate	Initial weights	2003 adjustment method	2006 adjustment method
Australia	9	81.6	9.3	8.7	9.2
	10	86.9	70.5	71.5	70.8
	11	83.9	20.0	19.7	19.8
Belgium	8	81.8	4.6	4.3	4.4
	9	90.7	30.5	30.8	31.1
	10	95.5	63.1	63.4	63.2
	11	87.4	1.1	1.0	1.0
Brazil	7	84.7	12.0	11.4	11.6
	8	89.2	21.6	21.6	22.0
	9	92.0	47.9	48.4	47.8
	10	92.9	17.9	18.1	18.0
Denmark	8	84.9	12.0	11.5	12.0
	9	89.8	85.3	85.9	85.3
Finland	8	88.6	11.5	11.0	11.7
	9	93.5	88.1	88.9	88.1
France	8	83.9	5.2	4.9	5.2
	9	90.5	34.8	35.4	34.8
	10	88.8	57.4	57.2	57.5
	11	88.4	2.5	2.4	2.4
Germany	8	85.8	12.5	11.9	11.9
	9	94.1	53.7	54.9	54.5
	10	93.2	28.2	27.8	28.2
	Vocational	74.0	3.6	3.6	3.6
Hong Kong – China	7	87.5	2.4	2.4	2.4
	8	86.3	9.6	9.2	9.3
	9	90.1	24.8	24.8	25.2
	10	92.9	63.0	63.5	63.0
Hungary	8	92.4	5.5	5.5	5.5
	9	93.4	65.7	66.1	65.7
	10	93.0	26.6	26.1	26.6
Ireland	8	65.9	3.4	2.8	2.7
	9	85.8	57.9	59.0	58.5
	10	83.1	21.3	21.0	21.2
	11	81.4	17.4	17.2	17.5
Italy	8	82.8	1.5	1.5	1.5
	9	83.1	15.4	14.1	15.0
	10	94.2	79.8	81.3	80.4
	11	87.8	3.0	2.8	2.8
Latvia	7	87.9	2.8	2.5	2.6
	8	94.2	16.2	15.8	16.3
	9	97.6	77.5	78.3	77.7
	10	93.0	3.1	3.0	3.0

(continued)

Table 6.2 (continued)

Country	Grade	Response rate	Initial weights	2003 adjustment method	2006 adjustment method
Mexico	7	88.9	2.5	2.3	2.3
	8	93.5	8.1	8.0	8.1
	9	95.7	33.5	33.6	33.2
	10	95.5	47.9	48.0	48.5
	11	98.5	5.0	5.2	5.1
	12	99.9	2.0	2.1	2.0
	96	75.4	1.0	0.9	0.8
Netherlands	8	86.3	3.9	3.8	3.7
	9	90.3	44.8	45.3	44.9
	10	89.9	50.7	50.4	50.7
Poland	8	77.6	3.7	3.2	3.8
	9	92.5	95.0	95.8	95.0
	10	95.1	0.6	0.6	0.6
Portugal	7	76.0	6.9	6.2	6.4
	8	82.2	12.8	12.3	12.8
	9	86.8	28.4	29.0	28.9
	10	88.9	49.4	50.2	49.6
	Vocational	90.6	2.1	2.1	2.1
Spain	8	64.1	8.6	6.7	7.0
	9	80.7	31.6	31.4	33.0
	10	96.2	59.6	61.8	59.8
Sweden	8	66.3	2.6	1.8	1.9
	9	92.2	95.2	95.9	95.9
	10	86.4	2.2	2.2	2.2
United States	9	84.1	10.7	10.0	10.7
	10	92.5	70.3	71.5	70.9
	11	88.2	17.1	16.7	16.5
Uruguay	7	66.7	8.2	7.2	7.5
	8	81.9	9.8	9.7	9.8
	9	86.4	17.1	17.3	17.3
	10	90.4	58.4	59.5	58.9
	11	85.0	6.5	6.2	6.6

There are seven countries where the difference between the initial weighted estimate and the 2003 method adjusted estimate is at least 1.0%: Australia grade 10, Germany grade 9, Ireland grade 9, Italy grades 9 and 10, Spain grades 8 and 10, the United States grade 10, and Uruguay grades 7 and 10. In each of these cases, the 2006 adjustment provides estimates closer to the initial-weight estimates than the 2003 procedure does. However, in the case of Spain, while improving the matches of the percentages at grades 8 and 10, the 2006 procedure actually increases the discrepancy at grade 9. This is because the 2006 procedure did not distinguish between grades 8 and 9. Thus in correcting the percentage in grade 10, the method overestimates the percentage in grade 9, at the expense of grade 8. Thus it is likely that some upwards nonresponse bias remains in this case, since presumably grade 9 students perform better than those in grade 8.

For most countries the grade distribution is very well preserved by the method used in 2006. However, in addition to Spain, there are several countries where there is still a small but noticeable distortion in the grade distributions with the 2006 method. This suggests that in these countries, even though the 2006 method is likely subject to less nonresponse bias than the 2003 method, some bias is likely to remain in the 2006 data. In Germany, Ireland, Portugal, and Sweden, the discrepancy in grade distribution between the distribution from the 2006 method, and that of the initial weights, suggests the likelihood of a small amount of upwards bias in the 2006 achievement means. For the United States this bias is likely to be downwards. But again, the differences discussed here are very small, and so would contribute little bias. For example, in Germany the initial weight distribution shows 53.7% of students in grade 9, whereas after nonresponse adjustments the percentage in the responding sample is 54.5%. In the United States the initial distribution shows 17.1% of students in grade 11, while after nonresponse adjustments the percentage in the responding sample is 16.5%.

Table 6.3 presents the mean estimate and its respective standard error in Mathematics, Reading, and Science, respectively, by country. The results for four countries differ most noticeably between the two weight adjustment methods, across the three assessment domains: Ireland, Poland, Portugal and Spain. These differences are highlighted in the table.

In Poland, Portugal and Spain, the student response rates vary both by gender and grade. The 2006 adjustment procedure leads to a lower country average performance in the three domains than the 2003 adjustment procedure does, as it more effectively addresses the nonresponse bias due to the differential response rates by grade and gender.

In Ireland, the better estimates for the grade distribution are responsible for the shift in the country average performance, since the response rates do not differ by gender. In this case the differences in results for the two procedures are in the opposite direction from those of the three countries above: the 2006 procedure gives higher mean scores for Ireland than the 2003 procedure. This is because the 2003 procedure gives slightly too much weight to grade 9 students, at the expense of grades 8, 10, and 11. The 2006 procedure redresses this underrepresentation at grades 10 and 11, reducing the weight for grade 9. However, as mentioned earlier, since the 2006 procedure still leaves grade 8 somewhat underrepresented it seems quite likely that the results for Ireland for 2006 actually have some upwards nonresponse bias.

In the case of Spain, the 2006 procedure has restored the correct relative weighting to boys and girls but, as mentioned, has tended to over represent grade 9 at the expense of grade 8. Thus even though the mean achievement results from the 2006 procedure are lower than those from the 2003 procedure, it seems likely that some upwards nonresponse bias remains for Spain even with the 2006 procedure.

Table 6.3 Country mean estimates in Mathematics, Reading and Science for 2006, based on the 2003-Method Adjusted Weights and the 2006-Method Adjusted Weights

	Mathematics					Reading					Science				
	2003-method adjusted weights		2006-method adjusted weights		Difference of means	2003-method adjusted weights		2006-method adjusted weights		Difference of means	2003-method adjusted weights		2006-method adjusted weights		Difference of means
	Mean	SE	Mean	SE		Mean	SE	Mean	SE		Mean	SE	Mean	SE	
Australia	520.0	(2.2)	519.9	(2.2)	0.1	512.9	(2.1)	512.9	(2.1)	0.0	527.0	(2.2)	526.9	(2.3)	0.1
Austria	505.4	(3.7)	505.5	(3.7)	–0.1	490.3	(4.1)	490.2	(4.1)	0.1	510.9	(3.9)	510.8	(3.9)	0.0
Belgium	520.5	(2.9)	520.3	(3.0)	0.1	500.9	(3.0)	500.9	(3.0)	0.0	510.4	(2.5)	510.4	(2.5)	0.0
Brazil	369.7	(2.9)	369.5	(2.9)	0.2	393.1	(3.7)	392.9	(3.7)	0.2	390.4	(2.8)	390.3	(2.8)	0.1
Canada	527.0	(2.1)	527.0	(2.0)	0.0	526.5	(2.6)	527.0	(2.4)	–0.5	534.1	(2.2)	534.5	(2.0)	–0.4
Czech Republic	509.9	(3.6)	509.9	(3.6)	0.1	482.9	(4.1)	482.7	(4.2)	0.2	512.9	(3.5)	512.9	(3.5)	0.1
Denmark	513.1	(2.6)	513.0	(2.6)	0.1	495.0	(3.1)	494.5	(3.2)	0.5	496.0	(3.1)	495.9	(3.1)	0.1
Finland	548.6	(2.3)	548.4	(2.3)	0.2	547.2	(2.2)	546.9	(2.1)	0.4	563.6	(2.0)	563.3	(2.0)	0.2
France	495.7	(3.2)	495.5	(3.2)	0.1	487.8	(4.0)	487.7	(4.1)	0.1	495.3	(3.4)	495.2	(3.4)	0.1
Germany	503.2	(3.9)	503.8	(3.9)	–0.6	494.6	(4.4)	494.9	(4.4)	–0.3	515.1	(3.8)	515.6	(3.8)	–0.5
Greece	459.0	(3.0)	459.2	(3.0)	–0.3	459.3	(4.1)	459.7	(4.0)	–0.4	473.0	(3.3)	473.4	(3.2)	–0.4
Hong Kong – China	547.4	(2.6)	547.5	(2.7)	–0.1	536.0	(2.4)	536.1	(2.4)	–0.1	542.2	(2.5)	542.2	(2.5)	0.0
Hungary	490.6	(2.9)	490.9	(2.9)	–0.3	482.2	(3.3)	482.4	(3.3)	–0.2	503.7	(2.7)	503.9	(2.7)	–0.3
Iceland	505.9	(1.8)	505.5	(1.8)	0.3	484.9	(1.9)	484.4	(1.9)	0.4	491.0	(1.7)	490.8	(1.6)	0.2
Indonesia	390.9	(5.6)	391.0	(5.6)	–0.1	393.0	(5.9)	392.9	(5.9)	0.0	393.4	(5.7)	393.5	(5.7)	–0.1
Ireland	500.6	(2.8)	501.5	(2.8)	–0.9	516.4	(3.6)	517.3	(3.5)	–1.0	507.4	(3.2)	508.3	(3.2)	–1.0
Italy	461.3	(2.3)	461.7	(2.3)	–0.4	468.5	(2.4)	468.5	(2.4)	0.0	475.2	(2.0)	475.4	(2.0)	–0.2
Japan	523.1	(3.3)	523.1	(3.3)	0.0	498.0	(3.6)	498.0	(3.6)	0.0	531.4	(3.4)	531.4	(3.4)	0.0
Korea	547.5	(3.8)	547.5	(3.8)	0.0	556.1	(3.8)	556.0	(3.8)	0.0	522.2	(3.4)	522.1	(3.4)	0.0
Latvia	486.4	(3.0)	486.2	(3.0)	0.3	480.0	(3.7)	479.5	(3.7)	0.5	489.9	(3.0)	489.5	(3.0)	0.3
Liechtenstein	525.3	(4.1)	525.0	(4.2)	0.4	511.0	(3.9)	510.4	(3.9)	0.6	522.6	(4.0)	522.2	(4.1)	0.4
Luxembourg	490.1	(1.1)	490.0	(1.1)	0.1	479.4	(1.3)	479.4	(1.3)	0.0	486.4	(1.1)	486.3	(1.1)	0.0

Macao	524.9	(1.3)	525.0	(1.3)	–0.1	492.3	(1.1)	492.3	(1.1)	0.0	510.8	(1.1)	510.8	(1.1)	0.0
Mexico	405.2	(3.0)	405.7	(2.9)	–0.4	410.2	(3.2)	410.5	(3.1)	–0.3	409.4	(2.8)	409.7	(2.7)	–0.2
Netherlands	530.3	(2.6)	530.7	(2.6)	–0.4	506.0	(3.0)	506.7	(2.9)	–0.7	524.5	(2.8)	524.9	(2.7)	–0.4
New Zealand	522.1	(2.4)	522.0	(2.4)	0.1	521.2	(3.0)	521.0	(3.0)	0.2	530.6	(2.7)	530.4	(2.7)	0.2
Norway	489.6	(2.6)	489.8	(2.6)	–0.3	484.3	(3.1)	484.3	(3.2)	0.0	486.4	(3.1)	486.5	(3.1)	–0.2
Poland	496.3	(2.4)	495.4	(2.4)	0.8	509.0	(2.8)	507.6	(2.8)	1.4	498.7	(2.3)	497.8	(2.3)	0.9
Portugal	467.2	(3.1)	466.2	(3.1)	1.0	473.5	(3.6)	472.3	(3.6)	1.2	475.5	(3.0)	474.3	(3.0)	1.2
Russia Federation	475.7	(3.9)	475.7	(3.9)	0.1	439.9	(4.4)	439.9	(4.3)	0.1	479.4	(3.7)	479.5	(3.7)	0.0
Serbia	435.3	(3.5)	435.4	(3.5)	–0.1	400.9	(3.5)	401.0	(3.5)	–0.2	435.6	(3.0)	435.6	(3.0)	–0.1
Slovak Republic	492.1	(2.8)	492.1	(2.8)	0.0	466.4	(3.1)	466.3	(3.1)	0.0	488.4	(2.6)	488.4	(2.6)	0.0
Spain	481.5	(2.3)	480.0	(2.3)	1.6	462.4	(2.2)	460.8	(2.2)	1.6	490.0	(2.6)	488.4	(2.6)	1.5
Sweden	502.2	(2.4)	502.4	(2.4)	–0.1	507.1	(3.5)	507.3	(3.4)	–0.2	503.2	(2.4)	503.3	(2.4)	–0.2
Switzerland	529.5	(3.1)	529.7	(3.2)	–0.1	499.3	(3.1)	499.3	(3.1)	0.0	511.4	(3.2)	511.5	(3.2)	–0.2
Thailand	417.0	(2.3)	417.1	(2.3)	–0.1	416.8	(2.6)	416.8	(2.6)	0.1	421.0	(2.2)	421.0	(2.1)	0.0
Tunisia	365.3	(4.0)	365.5	(4.0)	–0.2	380.5	(4.0)	380.3	(4.0)	0.2	385.3	(3.0)	385.5	(3.0)	–0.2
Turkey	423.8	(4.9)	423.9	(4.9)	–0.1	447.2	(4.2)	447.1	(4.2)	0.1	423.8	(3.8)	423.8	(3.8)	–0.1
United Kingdom	495.3	(2.1)	495.4	(2.1)	–0.1	494.8	(2.2)	495.1	(2.3)	–0.3	514.6	(2.3)	514.8	(2.3)	–0.2
United States[a]	474.9	(4.0)	474.4	(4.0)	0.5	NA	NA	NA	NA	NA	489.5	(4.2)	488.9	(4.2)	0.6
Uruguay	427.2	(2.6)	426.8	(2.6)	0.4	412.8	(3.4)	412.5	(3.4)	0.3	428.6	(2.8)	428.1	(2.7)	0.5

[a] No Reading results are available for the United States

In the cases of Poland and Portugal, the 2006 procedure has been effective in providing the appropriate distributions both by gender and by grade. Thus one can be optimistic that there is very little student nonresponse bias in the results for these two countries, using the 2006 weighting procedure. In all cases, across the three assessment domains, the differences in means from the two weighting procedures are less than one standard error, and only in the case of Spain do they exceed 0.5 standard errors.

In discussing the results shown in Tables 6.1 and 6.2, a total of twelve countries were noted as having potential biases remaining in their data as a result of the use of the 2003 nonresponse adjustment procedure. Yet the above discussion of Table 6.3 notes that an appreciable impact was seen in only four of those countries. Some explanation of this is in order.

In Austria, Denmark, Iceland, Italy, and Tunisia, it was noted that there was differential response for girls and boys (Table 6.1). It was already noted that, in the case of Austria, this translated into only a small bias in the gender distribution of the weights adjusted using the 2003 method. Thus it is not surprising that the results in Table 6.3 show almost no differences between the two methods in the results for Austria. In the case of Denmark, there is some difference evident in the results for Reading, with the 2006 method giving a mean score that is 0.5 points lower than the 2003 method. But the difference is much less for the other two domains (0.1 points in each case). This is a reflection of the fact that reading is more highly correlated with gender than are Mathematics and Science in Denmark (as in many other countries). In Iceland and Tunisia the differences between methods in mean scores are noticeable but small, ranging from 0.2 to 0.4 points. We defer the discussion of Italy until later.

In Australia, Germany, Italy, the United States, and Uruguay, differences in grade distribution were noted between the two methods (Table 6.2). In Australia the differences in mean scores between the two methods were 0.1 or 0.0 across the three domains. Although the data in Table 6.2 suggest that the 2006 method might be expected to result in somewhat lower mean scores, this appears to have been counteracted by the fact that the response rate was higher for boys than for girls, which is unusual in PISA. In Germany, the United States, and Uruguay, the 2006 method did result in mean scores that differed from those from the 2003 method by 0.3 to 0.6 points. As one might expect based on the results in Table 6.2, the 2006 method gives higher mean scores in Germany, and lower scores in the United States and Uruguay. But the sizes of the effects were smaller than in the four countries with the most noticeable effects, where the differences were typically 1.0 or greater.

Italy, like Spain, showed differences between the two methods in terms of both gender and grade distribution. The differences in mean scores between the two methods varied by subject. For reading, there was no difference, whereas for mathematics the mean score was 0.4 points higher using the 2006 method, while for science it was 0.2 points higher. Thus while the effects of the different methods on the gender and grade distributions were similar in nature in Italy and Spain, but more marked in Spain, the effect on the mean results for the two countries were quite different, since in Spain the 2006 method gave mean scores that were 1.5–1.6 points lower than the 2003 method. It is in fact somewhat surprising that, for Italy, the 2006 method, which gave rise to both more boys, and more grade 9 and fewer grade 10 students than the 2003 method, should have given rise to an identical mean reading score. This underscores the fact that the nonresponse adjustment procedures differed by more than just the use of gender and grade. The role of school varied across the two procedures, and this means that a simple decomposition into gender and grade effects does not explain the differences in results in all cases.

6.2 Conclusions

In summary, the new method of student nonresponse adjustment used in 2006 appears to be more efficient in reducing the potential bias due to the differential participation rates between boys and girls, and also generally comes closer for the distribution by grade to that of the initial weighted estimates. At the same time, the change in method does not appear to have generated any spurious changes in achievement means of any consequence. As the 2006 procedure clearly appears to reduce the potential for student nonresponse bias, it was adopted for PISA 2009 also. This means that there is no potential issue of trend artifacts due to changing nonresponse adjustment procedures between 2006 and 2009.

For those cases noted above where mean achievement changed by more than 1 point it seems clear that there is noticeable nonresponse bias, and that the 2006 method addressed this more effectively than the 2003 method. However, even in these cases the differences in results between the two methods, across 41 countries and three assessment domains, was always less than one standard error, and exceeded 0.5 standard errors only in the case of Spain.

The evidence from this analysis suggests that there remains some potential for small amounts of bias in the 2006 results due to differential student nonresponse. There is a limit to the extent to which weight adjustments are able to eliminate nonresponse bias. This points to the continuing need to eliminate differential nonresponse to the extent possible. Reductions in differential nonresponse across all student characteristics (not just gender and grade) can only be achieved realistically by extending the efforts to raise student response rates across the board.

References

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research, 5*, 215–238.

Kalton, G. (1983). *Compensating for missing survey data*. Ann Arbor, MI: Institute for Social Research, University of Michigan.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics, 19*, 81–97.

Kish, L. (1992). Weighting for unequal Pi. *Journal of Official Statistics, 8*, 183–200.

Little, R. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review, 54*, 139–157.

Organisation for Economic Co-operation and Development. (2002). PISA 2000 Technical Report. Paris: OECD.

Organisation for Economic Co-operation and Development. (2005). PISA 2003 Technical Report. Paris: OECD.

Organisation for Economic Co-operation and Development. (2009). PISA 2006 Technical Report. Paris: OECD.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research, 5*, 239–261.

Chapter 7
Reporting Differentiated Literacy Results in PISA by Using Multidimensional Adaptive Testing

Andreas Frey, Nicki-Nils Seitz, and Ulf Kröhne

Abstract Multidimensional adaptive testing (MAT) allows for substantial increases in measurement efficiency. It was examined whether this capability can be used to report reliable results for all 10 subdimensions of students' literacy in reading, mathematics and science considered in PISA. The responses of $N = 14{,}624$ students who participated in the PISA assessments of the years 2000, 2003 and 2006 in Germany were used to simulate unrestricted MAT, MAT with the multidimensional maximum priority index method (MMPI), and MAT with MMPI taking typical restrictions of the PISA assessments (treatment of link items, treatment of open items, grouping of items to units) into account. For MAT with MMPI the reliability coefficients for all subdimensions were larger than .80, as opposed to sequential testing based on the booklet design of PISA 2006. These advantages slightly lessened with the incorporation of PISA-typical restrictions. The findings demonstrate that MAT with MMPI can successfully be used for subdimensional reporting in PISA.

Keywords Computerized adaptive testing • Multidimensional adaptive testing • Item response theory • Programme for International Student Assessment • Large-scale assessments

A. Frey (✉) • N.-N. Seitz
Institute of Educational Science, Friedrich-Schiller-University Jena,
Am Planetarium 4, D-07743 Jena, Germany
e-mail: andreas.frey@uni-jena.de; nicki-nils.seitz@uni-jena.de

U. Kröhne
Center for Research on Educational Quality and Evaluation, German Institute
for International Educational Research, Schloßstraße 29, D-60486
Frankfurt am Main, Germany
e-mail: kroehne@dipf.de

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_7,

The Programme for International Student Assessment (PISA) is an international large-scale assessment of student achievement jointly developed by OECD member countries (http://www.pisa.oecd.org). The objective of PISA is to assess the degree to which 15-year-old students have acquired skills and knowledge essential for successful participation in the modern knowledge society. The first PISA assessment took place in the year 2000, with subsequent assessments conducted every 3 years. The study focuses on measuring students' literacy in reading, mathematics and science (cf. OECD, 2009a). In every assessment, one of the three domains is treated as the major domain (PISA 2000: reading, PISA 2003: mathematics, PISA 2006: science, PISA 2009: reading). For the major domain, differentiated results on subdimensions are reported, while for the other two domains, only results for one general dimension are reported. PISA results have received a lot of attention and have often stimulated intense and productive discussions about the effectiveness of educational systems. However, the valuable results come at a rather high price since large sample sizes of around 4,500–10,000 students are tested for each assessment in each country. Moreover, the tests are rather time-consuming and require 120 min of testing time per student for the cognitive items and an overall testing time of about 220 min per student.

All in all, the testing load associated with PISA is high and results in high costs. In the long run—especially if other large-scale assessments and tests are carried out at the same schools within a short time period—the willingness of schools and teachers to participate in PISA may decrease. For the students, long testing sessions may have a negative impact on their test-taking motivation. Thus, to ensure the cooperation of schools, teachers and students in the long term as well as to limit costs, it is beneficial to increase the efficiency of the testing procedures while maintaining the high level of precision. One testing procedure to substantially increase measurement efficiency of PISA without jeopardizing measurement precision lies in multidimensional adaptive testing (MAT; e.g., Frey & Seitz, 2009). The magnitude of possible increases in measurement efficiency by MAT was examined within a simulation study based on student responses in the PISA 2006 assessment in Germany by Frey and Seitz (2011). They found that measurement efficiency can be optimally increased 74% by using MAT instead of sequential testing with a fixed number of items in a fixed order (FIT) which is the current method of data collection in PISA. Additionally, they report that the average number of presented items can be reduced by MAT from 55 to 26 without a loss in measurement precision. Thus, if MAT is used instead of FIT, less than half of the items need to be presented. Nevertheless, the assessment of PISA is restricted in several ways. For example, the PISA item pool includes numerous items in the open response or short response formats. Many of these items cannot automatically be scored by a computer and are therefore difficult to use within MAT. If the incorporation of items in open response or short response format and other restrictions given in PISA are taken into account, Frey and Seitz report a measurement efficiency for MAT that is 40% higher compared to FIT.

The gain in measurement efficiency of MAT compared to FIT could be used to (a) reduce the number of items that need to be presented to participants keeping measurement precision constant, or (b) to increase measurement precision keeping the number of presented items constant. In the case of PISA, option (b) may increase the measurement precision to a level that allows reporting for all ten subdimensions of students' literacy and not only for the major domain of one assessment. Thereby, more differentiated results would be obtained within the same testing time. Nevertheless, if item selection is solely based on an optimality criterion within an unrestricted MAT algorithm, the number of presented items, and, in turn, reliability, may vary in an undesirable way between the subdimensions depending on item pool characteristics. While reliability may be unnecessarily high for some dimensions, for others it may be too low for reporting. Thus, a method has to be used in conjunction with MAT aligning reliability of all measured subdimensions. This can be achieved with the multidimensional generalization of the maximum priority index method (MMPI) recently proposed by Frey, Cheng, and Seitz (2010).

The present real data simulation study examines whether MAT used in conjunction with the MMPI method can produce sufficiently reliable results for all 10 subdimensions measured in PISA without presenting more items than in the regular assessments, even when taking PISA-specific restrictions into account. The text is organized as follows: First, the concept of multidimensional adaptive testing and the MMPI method are described. Subsequently, typical restrictions associated with the PISA assessments are depicted and the research questions stated. Then, the method and results of the simulation study are presented. Finally, the implications of the results regarding the use of multidimensional adaptive testing in PISA are discussed.

7.1 Multidimensional Adaptive Testing

Computerized Adaptive Testing (CAT) is a special approach to assess latent abilities in which the selection of the test items presented to an examinee is based on responses given by the examinee to previously administered items. The approach was originally formulated to measure one single dimension at a time and therefore, in the following is referred to as unidimensional computerized adaptive testing (UCAT). The aim of UCAT is to select items from an item pool for presentation that provide maximum information regarding the unidimensional ability level of the examinee. This results in a substantial increase in measurement efficiency. Compared to FIT, the number of items can typically be reduced by approximately half without a loss in measurement precision (e.g., Frey, 2012; Segall, 2005), when UCAT is used.

A natural generalization of UCAT is multidimensional adaptive testing (MAT). In MAT, several dimensions are assessed simultaneously using multidimensional item response theory (MIRT; e.g., Reckase, 2009) models as measurement models. A general form of a MIRT model is the multidimensional three parameter logistic test model (M3PL). The M3PL specifies the probability of a person j to correctly answer item i as a function of P latent abilities, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$, an 1 × P item discrimination vector $\mathbf{a}_i'$, an item difficulty parameter b_i, an item-specific pseudo-guessing parameter c_i, and an P × 1-vector **1**, consisting of 1's expanding the item difficulty to the multidimensional space:

$$P(U_{ij} = 1 \,|\, \boldsymbol{\theta}_j, \mathbf{a}_i', b_i, c_i) = c_i + (1 - c_i)\,\frac{e^{\mathbf{a}_i'}\,(\boldsymbol{\theta}_j - b_i\mathbf{1})}{1 + \,e^{\mathbf{a}_i'}\,(\boldsymbol{\theta}_j - b_i\mathbf{1})} \tag{7.1}$$

The elements of the item discrimination vector $\mathbf{a}_i'$ can be used to define whether an item loads on a dimension or not. If for every item the value 1 is assigned exactly once, the model describes between-item multidimensionality. Thus, the modeled dimensions are measured by a distinct set of items each. If for one or more items the value 1 is assigned more than once, the model describes within-item multidimensionality. In this case, the modeled dimensions are measured by overlapping item sets. The M2PL with between-item multidimensionality is very useful for large-scale assessments of student achievement as it allows for unequivocal interpretations of the measured dimensions. Although many other MIRT models can be used for MAT besides those mentioned, applications of more complex models are still rare. One exception describing the use of a hierarchical MIRT model with three levels comprising seven dimensions in a MAT framework is given by Segall (2001).

The major MAT approaches were introduced by Segall (1996), who describes a Bayesian as well as a maximum likelihood approach, and by van der Linden (1999), who uses maximum likelihood for item selection and ability estimation. In both, the item parameters are assumed to be known. The Bayesian approach of Segall (1996) is especially appealing since even higher measurement efficiency than in UCAT is achieved if correlated dimensions are measured. The increase in measurement efficiency is caused by using the multivariate prior distribution of the measured dimensions to optimize both the estimation of the latent ability vector, and the item selection process. For the estimation of the latent abilities, Segall (1996) proposes a multidimensional Bayes modal estimate using Fisher scoring and the variance-covariance matrix $\mathbf{\Phi}$. Regarding item selection, he suggests selecting the item from the item pool for presentation which maximizes the quantity

$$\left|\mathbf{W}_{t+i^*}\right| = \left|\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j) + \mathbf{I}(\boldsymbol{\theta}, u_{i^*}) + \mathbf{\Phi}^{-1}\right| \tag{7.2}$$

Thus, the item i^* is selected, which results in the largest determinant of the matrix $\mathbf{W}_{t+i^*}$, which is based on the information matrix of the previously t administered items, $\mathbf{I}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_j)$, the information matrix of a response u_{i*} to item i^*, $\mathbf{I}(\boldsymbol{\theta}, u_{i^*})$, and the inverse of the variance-covariance matrix of the prior distribution of the measured dimensions $\mathbf{\Phi}^{-1}$. This item provides the largest decrement in the volume of the credibility ellipsoid around the vector of latent abilities $\hat{\boldsymbol{\theta}}_j$.

The sketched Bayesian MAT approach showed very high measurement efficiency in simulation studies when several correlated dimensions were measured (Frey & Seitz, 2010; Segall, 1996; Wang & Chen, 2004;). For typical large-scale assessment situations with three to five highly correlated dimensions, the mean number of items that need to be presented to the participant can be reduced by 30–50%, if MAT is used instead of UCAT.

7.2 Multidimensional Maximum Priority Index Method

The maximum priority index method (MPI) was proposed for UCAT to accommodate for flexible content balancing by Cheng, Chang, Douglas, and Guo (2009) as well as for other constraints such as exposure control (Cheng & Chang, 2009).

The generalization of the method to the multidimensional case was proposed under the abbreviation MMPI by Frey et al. (2010).

MPI as well as MMPI strive to maximize a priority index. The priority index for each item indicates its desirability in terms of both statistical property as well as the urgency of having it to fulfill non-statistical constraints. The MMPI is based on an $I \times P$ constraint relevance matrix **C**, which content matter experts must specify before the assessment. Within **C**, *I* indicates the number of items in the pool and *P* the number of dimensions. $c_{i^* p} = 1$ if item $i = 1, \ldots, I$ loads on dimension $p = 1, \ldots, P$; otherwise $c_{i^* p} = 0$. The priority index (PI) for the Bayesian MAT approach of Segall (1996) is defined as:

$$\mathrm{PI} = \left| \mathbf{W}_{t+i^*} \right| \prod_{p=1}^{P} f_p^{c_{i^* p}} \tag{7.3}$$

Thus, the Bayesian item selection criterion from Eq. 7.2 is multiplied with a term quantifying the desirability of the candidate item i^*. f_p is given by the difference of the number of required items T_p and the number of presented items t_p divided by T_p. That is:

$$f_p = \frac{\left(T_p - t_p\right)}{T_p} \tag{7.4}$$

This ratio is 1 when no item has been selected from dimension *p*, and it gets smaller and smaller as more items from this dimension are picked, until it reaches 0, when the required number of items from dimension *p* has been given to examinees.

By multiplying $\prod_{p=1}^{P} f_p^{c_{i^* p}}$ with the determinant of the Bayesian information matrix $\mathbf{W}_{t+i^*}$, the MMPI method tries to strike a balance between fulfilling the content coverage requirement and statistical optimality. The item with the largest PI is selected as the next item to be presented to the test taker. If an equal number of items is requested for all dimensions, application of this procedure will align the reliability of the measured dimensions if the multidimensional Raschmodel and comparable item pools are used.

7.3 Restrictions in PISA

When considering using MAT to assess students' literacy in PISA, one must take into consideration that some features of PISA are not optimal for using MAT. The first feature is the PISA item pool: The entailed number of items is comparatively small, the majority of items have a medium item difficulty and only some items have very high or very low item difficulties. Thus, MAT based on the PISA items, will work below its optimal performance level in cases of extreme abilities. Second, PISA has a couple of restrictions. The most prominent restrictions needing consideration, when using MAT instead of FIT are:

- The item pool used in the PISA assessments from 2000 to 2006 contains 49% items in *open response* or *short response* format. Many items of these types cannot be directly scored by a computer. Hence, the responses given cannot be used within an adaptive testing procedure to revise the provisional ability vector.
- More than half of the items used in PISA 2006 (54%) are so-called *link items* which had already been used in previous PISA assessments. By use of an appropriate booklet design, the link items are presented to the student sample with a fixed relative frequency, allowing linking of assessments from different years. An unrestricted adaptive algorithm may not result in the desired relative frequencies of presented link items and, therefore, may jeopardize trend reporting.
- Only 13% of the items used in the PISA assessments from 2000 to 2006 are single items. All other items are grouped in so called units (testlets). Items of one unit are connected to the same stimulus. Splitting up units and selecting single items adaptively may result in the student being presented the same stimulus multiple times. This can be problematic regarding acceptance by the student and may invalidate item parameters.

7.4 Research Questions

MAT's capability to substantially increase measurement efficiency can theoretically be used to foster measurement precision of the 10 subdimensions considered in PISA. This would allow reporting for all subdimensions and not only for those connected to the major domain of the current assessment. Nevertheless, since the PISA item pool is not optimal for MAT, an unrestricted MAT algorithm may result in reliability coefficients varying greatly between the subdimensions. For some dimensions, the reliability may be too low to allow reporting results with sufficient precision. The MMPI strives to solve this problem. By aligning the number of presented items among all measured dimensions, it aims to produce a reliability that is sufficiently high for reporting for all subdimensions. Here, a reliability of .80 for a model without conditioning is considered to be sufficiently high. This value lies within the reliability range observed in the comparison of the 16 federal states of Germany in PISA 2006 (Prenzel et al., 2008) which varied from .78 (for reading in one federal state) to .93 (for science in one federal state) for scaling models without conditioning. Hence, if a minimum reliability of .80 is reached for all dimensions in the present study, their reliability is at least as high as the reliability of the scales used for reporting in the comparison of the German federal states in PISA 2006. Since the special version of the MMPI was not used in any previous study with the purpose of aligning reliability of several dimensions, it has to be shown whether it performs as expected.

To allow for direct interpretations of the results of the present study regarding PISA, restrictions typical for PISA should be considered. All restrictions mentioned in the previous section can be included in the MAT-algorithm. Nevertheless, the incorporation of these restrictions will decrease the measurement efficiency of MAT to a certain degree. It is not yet known whether or not MAT with MMPI will still allow producing scores for all 10 subdimensions with sufficient reliability. This study examines the following two research questions:

1. Is it possible to reach reliability coefficients larger than .80 for all 10 subdimensions considered in PISA by using MAT with MMPI?
2. Is it possible to reach reliability coefficients larger than .80 for all 10 subdimensions considered in PISA by using MAT with MMPI if the typical restrictions associated with PISA assessments are taken into account?

These research questions are answered by means of a real data simulation. To allow direct interpretations regarding an application of MAT for the assessment of students' literacy in PISA, the simulation design is specified to match the conditions of the PISA 2006 assessment as closely as possible.

7.5 Method

7.5.1 Sample

The study is based on the responses of 14,624 15-year-old students who participated in the PISA assessments during the years 2000 ($n = 5{,}073$), 2003 ($n = 4{,}660$), and 2006 ($n = 4{,}891$) in Germany. The answers of these students were used for the international PISA reports (OECD, 2001, 2004, 2007). Further descriptions of the samples can be found in the respective technical reports (Adams & Wu, 2002; OECD, 2005, 2009b). The responses were used to estimate the item parameters and the multidimensional ability distribution. Both were needed to simulate an application of MAT in PISA. Details can be found in Sect. 7.5.3.

7.5.2 Design

Four testing algorithms were compared with regards to reliability. The reference condition FIT was contrasted with MAT without restrictions, MAT with MMPI (MAT + MMPI) and MAT with MMPI, taking restrictions typical for the PISA assessments into account (MAT + MMPI + R).

As dependent variable, an index representing the reliability as a measurement design effect is calculated. In PISA a marginal item response theory model is used where estimates of student scores are not of interest and are therefore not produced.

An appropriate reliability index for marginal models was proposed by Adams (2005) with the EAP/PV reliability index. The EAP/PV reliability index is defined by the variance of the expected a-posteriori estimates var(EAP) and true population variance σ^2 (Adams, 2005):

$$\mathrm{Rel}_{\mathrm{EAP/PV}} = \frac{Var(\mathrm{EAP})}{\sigma^2} \tag{7.5}$$

The value of this reliability index is the average proportion of the uncertainty in the location of each student. This index was used in the present study without any conditioning. Thus, only the information stemming from the item responses is used for the calculation of the EAP/PV reliability index.

7.5.3 Procedure

The simulation was accomplished in three steps: the generation of item and person parameters, the generation of responses, and the actual simulation of the testing procedure. Details of the three steps are provided in the following sections.

7.5.3.1 Generation of Item and Person Parameters

The complete item pool of the present study consisted of all 348 items used in the assessments of PISA 2000, PISA 2003 and PISA 2006. The items are divided into 129 reading items, 95 mathematics items, and 124 science items. To obtain a common set of item parameters, the responses of the complete sample of 14,624 students were scaled with the Raschmodel for the dichotomously scored items and the partial credit model (Masters & Wright, 1997) for items with multiple score categories. In accordance with the international procedures of PISA, a separate unidimensional model was fitted for reading, mathematics and science using ACER ConQuest (Wu, Adams, Wilson, & Haldane, 2007). In the following, this initial scaling is referred to as *Scaling 1*. The resulting set of item parameters was used in all conditions.

In *Scaling 2*, the responses of the subsample of $n = 4{,}891$ students who enrolled in PISA 2006 in Germany were scaled with a 10-dimensional Raschmodel for the dichotomously scored items and partial credit model for the polytomously scored items. The 10 dimensions represented the following 10 subdimensions which are considered in PISA: *retrieving information* (READ 1), *interpreting texts* (READ 2), *reflection and evaluation* (READ 3), *space and shape* (MATH 1), *change and relationships* (MATH 2), *quantity* (MATH 3), *uncertainty* (MATH 4), *identifying scientific issues* (SCIE 1), *explaining phenomena scientifically* (SCIE 2), and *using scientific evidence* (SCIE 3). The item parameters were anchored at the values retrieved from Scaling 1. The resulting means and the variance-covariance matrix of the multidimensional latent distribution were used for the generation of the responses and the simulation of the testing procedure.

7.5.3.2 Generation of Responses

In MAT, theoretically, every item of the complete item pool could be presented to every student. Thus, for simulating MAT, a response is needed from every student to every item. Because each student who participated in one of the PISA assessments only answered about 4/13 of the item pool used in one assessment, which, in turn, is only part of the complete item pool, a complete response matrix was generated. Therefore, individual values in reading, mathematics and science were randomly drawn for 4,891 simulees from the multidimensional latent distribution derived from Scaling 2 under the assumption of multivariate normality. These person parameters were considered as true ability parameters $\boldsymbol{\theta}_j$. Together with the item parameters from Scaling 1, $\boldsymbol{\theta}_j$ was used to generate a response for each simulee to each item of the item pool based on the 10-dimensional Raschmodel. The mean item parameter was used for the items with multiple score categories. To account for the statistical uncertainty of the simulated answering process 100 replications were calculated. The resulting 100 complete response matrices of the size 4,891 × 348 served as the basis for the simulation of the testing procedure. The final statistics were derived by averaging the replications.

7.5.3.3 Simulation of the Testing Procedure

The actual testing procedure was simulated using the statistical package SAS 9.2. The following four testing conditions were specified.

Condition 1: Fixed Item Testing

Within the reference condition, FIT, the characteristics of the PISA 2006 assessment were rebuilt. The PISA 2006 booklet design (cf. Frey, Hartig, & Rupp, 2009; OECD, 2009b), comprising 13 booklets, each with a testing time of 120 min, was used to assign the items to the simulees. The item pool of the PISA 2006 assessment consisted of 179 items measuring reading, mathematics and science. Each item was included in one of 13 item clusters, which were systematically assigned to booklets and positions in booklets (Table 7.1).

In the present study the booklet design was used to select a set of responses from the complete response matrix for each simulee. The responses to all other items were treated as not administered for this simulee. To arrive at the final statistics for one replication, the selected responses were scaled with a 10-dimensional Raschmodel with the item parameters anchored at the values from Scaling 1.

Table 7.1 Booklet Design of PISA 2006

	Booklet												
Position	1	2	3	4	5	6	7	8	9	10	11	12	13
1	S1	S2	S3	S4	S5	S6	S7	M1	M2	M3	M4	R1	R2
2	S2	S3	S4	M3	S6	R2	R1	M2	S1	M4	S5	M1	S7
3	S4	M3	M4	S5	S7	R1	M2	S2	S3	S6	R2	S1	M1
4	S7	R1	M1	M2	S3	S4	M4	S6	R2	S1	S2	S5	M3

Note: R1–R2: reading clusters; M1–M4: mathematics clusters; S1–S7: science clusters

Condition 2: Unconstrained MAT

The booklet design was only used in the FIT condition. In all other conditions MAT was used. The first presented item was randomly chosen from the complete item pool. Adaptive item selection started with the second item. Item selection and provisional ability estimation were based on the Bayesian approach of Segall (1996), making use of the item parameters from Scaling 1, and the variance-covariance matrix of the prior distribution $\mathbf{\Phi}$ from Scaling 2. In the unconstrained MAT condition, Eq. 7.2 served as item selection criterion. The responses to the selected items were taken from the complete response matrix; not selected items were treated as not administered. The test was terminated when the next item would have exceeded the maximum testing time of 120 min. The calculation of the testing time was based on the testing time scheduled for item delivery in PISA 2006, which was 2.14 min for reading items, 2.50 min for mathematics items, and 2.05 min for science items. For the calculation of the final results for one replication, the selected responses were scaled with a 10-dimensional Raschmodel with the same specifications as in the condition FIT.

Condition 3: MAT with MMPI

The same procedure as described for unrestricted MAT was applied for condition 3 and extended by MMPI as described above and specified by Eqs. 7.3 and 7.4. The test length was set to 55 items. This was achieved by presenting $t_p = 5$ items for five subdimensions and $t_p = 6$ items to the other five subdimensions. To produce comparable overall numbers of required items per subdimension p, the dimensions measured with five or six items respectively were randomly selected for each simulee prior to the simulation of the testing procedure.

Condition 4: MAT with MMPI taking restrictions typical for PISA into account

In this condition, the procedure used in the MAT with MMPI condition was supplemented by taking the grouping of items to units, link items, and items that cannot be scored directly by a computer into account.

With regard to *units*, complete units were selected instead of single items. The size of the units contained in the complete PISA item pool ranges from one to seven items. The mean item difficulty of units with several items is predominantly around 0. Hence, the summed or averaged item information would not have been a good criterion for selecting units. To enable a better adjustment of the selected units to the provisional ability vector, the unit including the item with the highest information was selected for presentation.

Additionally, an algorithm was implemented to ensure that each participant received at least the number of *link items* that were given to the students on average in PISA 2006. To achieve this, beginning after the first randomly selected unit, unit selection was restricted to link units. Most link units contain only link items; some contain one non-link item. Only link units were considered as candidate units until a minimum of two link units for reading, six link units for mathematics, and four link units for science were selected. This was accomplished by using MMPI. After the desired numbers of link units were presented to the simulee, all remaining units were considered as candidate units.

Lastly, *items that cannot directly be scored by a computer* were also taken into account. A qualitative analysis revealed that 236 items of the complete item pool (68%) can directly be scored by a computer. A human coder must score the remaining 112 items (32%). Only the 236 items of the first group were used to revise the provisional ability vector $\widehat{\boldsymbol{\theta}}$. The other items were presented to the simulees as well, but the responses were only considered in the final scaling.

7.6 Results

First, the psychometric properties of the literacy scales in reading, mathematics and science will be described. Then, results will be presented to answer the research questions.

7.6.1 Scaling Outcomes

The characteristics of the item pool strongly influence the performance of adaptive tests (e.g., Veldkamp & van der Linden, 2010). The item parameters for the 10 subdimensions were taken from Scaling 1. Characteristics of the item sets are shown in Table 7.2.

Generally, the numbers of items assigned to the subdimensions are quite small. Furthermore, the sizes of the item sets vary between the subdimensions. While dimension MATH 4 contains only 20 items, the number of items is considerably larger for dimension READ 2 (59 items) and dimension SCIE 2 (57 items). The mean and the standard deviation of the item parameters also show variations with respect to the subdimensions. The highest mean item difficulty is observed for dimension MATH 4 (0.40); the lowest for dimension READ 1 (–0.67). Furthermore, the frequency of items with a medium difficulty is high while the frequency of items decreases towards the extremes of the ability scale for all subdimensions. The largest standard deviation of the item parameters can be found for MATH 2 (1.51), covering a range of more than 7 logits. SCIE 3 shows the smallest standard deviation of the item parameters (0.82), which are covering the range of about 4 logits.

Table 7.2 Item pool characteristics

Sub-dimension	Number of items	Item difficulty *M*	*SD*	Min	Max
READ 1 (retrieving information)	38	−0.67	1.17	−3.08	1.97
READ 2 (interpreting texts)	59	−0.65	0.90	−2.88	1.62
READ 3 (reflection and evaluation)	32	0.05	1.11	−2.19	2.62
MATH 1 (space and shape)	22	0.22	1.27	−1.62	2.14
MATH 2 (change and relationships)	29	0.02	1.51	−3.94	3.33
MATH 3 (quality)	24	−0.52	0.91	−2.76	0.90
MATH 4 (uncertainty)	20	0.40	0.97	−1.93	2.22
SCIE 1 (identifying scientific issues)	26	−0.38	0.92	−2.36	1.09
SCIE 2 (explaining phenomena scientifically)	57	−0.33	1.11	−3.02	2.13
SCIE 3 (using scientific evidence)	41	−0.15	0.82	−2.43	1.75

In summary, the item parameter distributions are far from optimal for MAT. More items per dimension and more items of extreme difficulty should be available for all subdimensions to allow an optimal functioning of MAT. Nevertheless, even though the item pool is not optimal, it can be used for MAT and will allow some adjustment of the presented items to the response behavior.

The population characteristics obtained from Scaling 2 are shown in Table 7.3. With values between 2.77 (READ 2) and 1.15 (SCIE 1), the variances (values on the main diagonal) are quite different for the subdimensions. The latent correlations between the subdimensions (below the main diagonal) have values between .67 and .95. This indicates a moderate to very high connection between the subdimensions. The covariances between the subdimensions are shown above the main diagonal. With the variances, these form the variance-covariance matrix $\mathbf{\Phi}$ used in all MAT conditions for item selection and person parameter estimation within the testing procedure.

7.6.2 Reliability of the Subdimensions

The first research question asks whether MAT with MMPI can be used to produce reliability estimates larger than .80 for all 10 subdimensions considered in PISA. The results given in Table 7.4 show that the reliability coefficients exceeded .80 for all subdimensions in the condition MAT + MMPI (Range: .83–.88) as intended. The same does not hold in the conditions FIT (Range: .69–.87) and unconstrained MAT (Range: .64–.91). For FIT, in six of ten subdimensions the reliability coefficients are smaller than .80. The reliability coefficients for the seven subdimensions for reading and mathematics are significantly smaller in the condition FIT than in the condition MAT + MMPI (Fig. 7.1). The scientific subdimensions show no significant differences between FIT and MAT + MMPI. The comparably high reliabilities for the scientific subdimensions in the condition FIT are achieved by administering a lot of items for these subdimensions; 58% of all presented items were science items.

Table 7.3 Population characteristics

	Dimension									
Subdimension	READ 1	READ 2	READ 3	MATH 1	MATH 2	MATH 3	MATH 4	SCIE 1	SCIE 2	SCIE 3
READ 1	1.79	2.06	1.45	1.38	1.58	1.34	1.33	1.27	1.31	1.46
READ 2	0.92	2.77	1.88	1.69	1.97	1.68	1.63	1.53	1.59	1.79
READ 3	0.87	0.90	1.57	1.14	1.35	1.14	1.14	1.07	1.12	1.28
MATH 1	0.76	0.74	0.67	1.86	1.67	1.54	1.50	1.12	1.23	1.35
MATH 2	0.84	0.84	0.77	0.87	1.97	1.70	1.59	1.28	1.37	1.52
MATH 3	0.76	0.77	0.69	0.86	0.92	1.73	1.48	1.13	1.21	1.33
MATH 4	0.80	0.79	0.73	0.88	0.91	0.90	1.56	1.09	1.18	1.29
SCIE 1	0.89	0.86	0.80	0.77	0.85	0.80	0.81	1.15	1.10	1.21
SCIE 2	0.88	0.86	0.80	0.81	0.88	0.82	0.85	0.92	1.24	1.28
SCIE 3	0.90	0.89	0.84	0.82	0.90	0.84	0.86	0.93	0.95	1.46

Note. Values on the main diagonal are variances, values below the main diagonal are correlations, values above the main diagonal are covariances

Table 7.4 Reliability coefficients per subdimension for four testing algorithms

	FIT		MAT		MAT + MMPI		MAT + MMPI + R	
Sub-dimension	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
READ 1	.76	0.02	.88	0.01	.87	0.01	.84	0.01
READ 2	.79	0.01	.91	0.01	.88	0.01	.87	0.01
READ 3	.69	0.02	.86	0.01	.83	0.01	.79	0.01
MATH 1	.74	0.02	.86	0.01	.84	0.01	.81	0.01
MATH 2	.81	0.01	.89	0.01	.88	0.01	.85	0.01
MATH 3	.76	0.01	.86	0.01	.84	0.01	.81	0.01
MATH 4	.76	0.01	.86	0.01	.85	0.01	.81	0.01
SCIE 1	.82	0.01	.80	0.02	.84	0.01	.80	0.02
SCIE 2	.85	0.01	.80	0.02	.85	0.01	.82	0.01
SCIE 3	.87	0.01	.64	0.04	.87	0.01	.86	0.01

Note. FIT: fixed item testing, MAT: multidimensional adaptive testing, MMPI: maximum priority index, R: restrictions

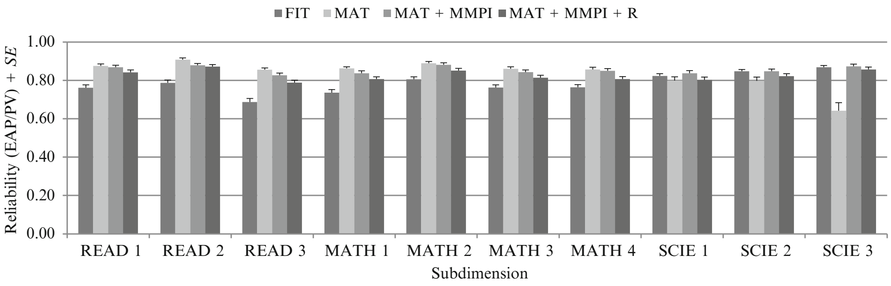


Fig. 7.1 Reliability coefficients per subdimension for four testing algorithms

In summary, the proposed new version of the MMPI largely performed as expected. Nevertheless, even if the variation of the reliability coefficients between dimensions was reduced by MMPI, they still vary somewhat and are not completely aligned to one another. Possible reasons for this result are taken up in Sect. 7.7.

The second research questions asks whether it is possible to reach reliability coefficients larger than .80 for all 10 subdimensions considered in PISA by MAT with MMPI if typical restrictions of PISA are taken into account. As can be seen in Table 7.4 and Fig. 7.1, MAT with MMPI and restrictions performed only slightly worse than MAT with MMPI. In the condition MAT + MMPI + R, only the reliability coefficient for READ 3 (.79) did not completely reach the desired value of .80. With a range of .79–.87, the reliability coefficients vary over an interval that is considerably smaller than for FIT but a bit larger than for MAT + MMPI. The good performance of MAT + MMPI + R is underlined by a mean reliability coefficient of .83 which is a non-negligible improvement compared to FIT (.78). Summarizing, the advantages of MAT with MMPI compared to FIT are slightly decreased.

7.7 Discussion

It was examined whether the high measurement efficiency of MAT can be used to produce reliable results for all 10 subdimensions of students' literacy in reading, mathematics and science considered in PISA. The results are promising for MAT. As intended, MAT used in conjunction with MMPI produced reliability coefficients larger than .80 for all subdimensions. In contrast to these findings, in six of ten subdimensions the reliability coefficients are smaller than .80 when FIT is used. If typical restrictions of the PISA assessments are taken into account, the advantage of MAT with MMPI compared to FIT is only slightly smaller. Nevertheless, the reliability coefficients are still larger than .80 for 9 out of 10 dimensions. One dimension missed the target reliability slightly (.79). The mean reliability was improved from .78 (FIT) to .83 by (MAT+MMPI+R). Note that in practice, the collected responses can surely be used to calculate score distributions for the major domains of reading, mathematics, and science as well as for the 10 subdimensions. Thus, by the use of MAT with MMPI the usual results can be reported plus additional information on all subdimensions of the literacy scales.

These promising findings may even be augmented. In the present simulation study the correlations between the measured dimensions are only used to optimize the item selection process. Further increases in the accuracy of population estimates can be expected if assumptions about the correlations between the measured dimensions drawn from previous assessments are also used for the estimation of the final ability estimates within a Bayesian framework. Nevertheless, this would imply rather strong assumptions that may not be realistic for PISA. Another possibility to further foster reliability is to use background information for conditioning within the item selection process. Thereby, not only the item response part of the mixed coefficients multinomial logit model (Adams, Wilson, & Wang, 1997) is used for MAT (as in the present study) but also the population model. When testing takes place, not all variables and indices used for conditioning in PISA in the background model are available. Nevertheless, several important variables, like the stratification variables and information stemming from the student tracking form, are. It is an open research question which decrease in the statistical uncertainty of the estimated ability distribution can be achieved if this information is used as conditioning variables within the item selection process.

Despite the promising findings mentioned above, the results also show that the variation in reliability coefficients between dimensions was reduced, but not completely removed by MMPI. This can be explained by the interdependence of the restrictions examined in this study in combination with the relatively small item sets available for each subdimension. The subdimension with the lowest reliability in the condition MAT+MMPI+R, READ 3, has a non-optimal combination of item specifications. First, all READ 3 items are included in units with items from the other reading subscales. Thus, if the adaptive algorithm strikes to optimize the measurement precision of READ 3, items for at least one of the other reading subscales are presented as well. Second, only eight out of the 32 items for this subdimension can be scored automatically by a computer. Since only items that can be scored directly are used to revise the provisional ability estimation within the test, the estimate for READ 3 stays relatively imprecise compared to subscales where the percentage of items that can be scored directly is larger. Consequently, for the cases where the provisional estimate for READ 3 is far off the true value, units are selected that do not provide maximum information. This results in a relatively low reliability.

When considering the usage of MAT in PISA, we propose precisely costing out all possible alternatives including, for instance, multi-stage testing. Obviously, one must consider the fact that the necessary hardware must be on standby and available at the testing location when computerized testing is used. It should be shown beforehand, that the high measurement efficiency of MAT and possible other advantages of a computer-based test delivery will outweigh the costs induced by MAT. Other general advantages and disadvantages of computer-based testing compared to paper-and-pencil testing have already been discussed in detail and are thus not repeated here (cf. Bartram & Hambleton, 2005; Kröhne & Martens, 2011; Parshall, Spray, Kalohn, & Davey, 2002). In particular, a combination of paper-and-pencil and computer-based testing should be considered when computerized assessment of the whole framework is not feasible in terms of content coverage. Moreover, the present simulation study only highlights the most important formal restrictions of PISA. Other possible restrictions as well as the psychological effects of the testing algorithm on students' response behavior are not modeled. These may also affect the multidimensional ability distribution. Since the differences interpreted in PISA are often rather small, systematic effects due to a change of the testing algorithm may lead to invalid inferences. Thus, whether a stable link to the previous assessments can be established is a challenging but also exciting empirical question.

In conclusion, the present real data simulation illustrates that MAT can be advantageous for reporting results for the literacy subdimensions in PISA even under constrained conditions. We suggest considering this highly efficient way of testing when obtaining differentiated results on all 10 subdimensions in every assessment of PISA is the aim. In particular, MAT will be a promising advancement for the assessments in PISA when testing should be switched from paper-and-pencil to computer based assessment and when item pools are revised and extended for computerized testing.

Acknowledgements The preparation of this article was supported by grant FR 2552/2-2 from the German Research Foundation (DFG) in the Priority Programme "Models of Competencies for Assessment of Individual Learning Outcomes and the Evaluation of Educational Processes" (SPP 1293) and from the Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162–172.

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

Adams, R. J., & Wu, M. (Eds.). (2002). *PISA 2000 technical report*. Paris: OECD.

Bartram, D., & Hambleton, R. K. (2005). *Computer-based testing and the internet: Issues and advances*. New York: Wiley.

Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*, 369–383.

Cheng, Y., Chang, H., Douglas, J., & Guo, F. (2009). Constraint-weighted a-stratification for computerized adaptive testing with non-statistical constraints: Balancing measurement efficiency and exposure control. *Educational and Psychological Measurement, 69*, 35–49.

Frey, A. (2012). Adaptives Testen [Adaptive testing]. In Moosbrugger, H. & Kelava, A. (Eds.), *Testtheorie und Fragebogenkonstruktion*, 2nd edn. [*Test theory and construction of questionnaires*, 2nd edn.] (pp. 261–278). Berlin/Heidelberg, Germany: Springer.

Frey, A., Cheng, Y., & Seitz, N. N. (2010, June). *Content balancing with the maximum priority index method in multidimensional adaptive testing*. Paper presented at the conference of the International Association for Computerized Adaptive Testing, Arnhem, the Netherlands.

Frey, A., Hartig, J., & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: theory and practice. *Educational Measurement: Issues and Practice, 28*, 39–53.

Frey, A., & Seitz, N. N. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation, 35*, 89–94.

Frey, A., & Seitz, N. N. (2010). Multidimensionale adaptive Kompetenzdiagnostik: Ergebnisse zur Messeffizienz [Multidimensional adaptive testing of competencies: Results regarding measurement efficiency]. *Zeitschrift für Pädagogik, Beiheft, 56*, 40–51.

Frey, A., & Seitz, N. N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in PISA. *Educational and Psychological Measurement, 71*, 503–522.

Kröhne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft, Sonderheft, 14*, 169–186.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–122). New York: Springer.

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris: OECD.

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

OECD. (2005). *PISA 2003 technical report*. Paris: OECD.

OECD. (2007). *PISA 2006. Science competencies for tomorrow`s world (Vol. 1: Analysis)*. Paris: OECD.

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

OECD. (2009b). *PISA 2006 technical report*. Paris: OECD.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York/Berlin/Heidelberg, Germany: Springer.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., & Pekrun, R. (Eds.). (2008). *PISA 2006 in Deutschland: Die Kompetenzen der Jugendlichen im dritten Ländervergleich [PISA 2006 in Germany: The third comparison of student competences]*. Münster: Waxmann.

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

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331–354.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika, 66*, 79–97.

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement*. New York: Academic.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics, 24*, 398–412.

Veldkamp, B. P., & van der Linden, W. J. (2010). Designing item pools for adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 231–245). New York: Springer.

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement, 28*, 450–480.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Melbourne, Australia: ACER Press.

Part III
Context Related Research

Introduction: Context Related Research on PISA

Eckhard Klieme

When International Large Scale Assessments (henceforth abbreviated as LSA) were invented in the late 1950s, the creation of international benchmarks for educational policy was not meant to be their prime goal. Rather, the founders and principal investigators of the International Association for the Evaluation of Educational Achievement (IEA), such as Benjamin Bloom, R.L. Thorndike, J.B. Carroll, Torsten Husen, and Neville Postleswaith, who all were renowned educational researchers, aimed to evaluate the intercultural validity of findings on the conditions of successful learning processes, which had up to then only been assessed in particular cultural contexts. Also, they intended to study effects on the system level, i.e. the impact of differently structured education systems on educational practices and student achievement, thus revolutionizing comparative educational research which so far had been arguing in a purely qualitative-historical style.

Today, International LSA – both the IEA studies and the OECD studies such as PISA – are perceived, and designed, as studies that primarily serve the needs of educational policymaking. However, there are still multiple stakes involved. The views endorsed by different stakeholders in the participating countries may be broken down into the following broad areas:

- LSA establish a monitoring structure that provides reliable comparative information on education systems, describing system structures as well as the functioning and the productivity (i.e. the gross outcome or "yield") of education systems. LSA data cover student career paths up to secondary level, school characteristics, school governance, student performance and motivation, as well as equity issues (such as performance by gender as well as socio-economic background).

E. Klieme (✉)
German Institute for International Educational Research (DIPF), Schloßstraße 29,
60486 Frankfurt am Main, Germany
e-mail: klieme@dipf.de

- LSA also contribute to our knowledge base on educational effectiveness. The studies observe patterns of relationships between inputs, processes and outcomes of education. Thus, they help to understand how educational outcomes are "produced". Firstly, LSA allow for a decomposition of variation of student performance by individual, school and system levels. Moreover, they provide data about multiple factors – covering these three levels – which, according to previous research, are expected to impact student performance in specific domains like reading, mathematics, or science. In addition to describing these factors, LSA allow to estimate their direct and indirect relationships to student performance and other outcomes. Statistical models, using multi-level LSA data, help to reconstruct and understand the complex relationships between input and process factors, and how they interact in "producing" student outcomes. If data on resources and costs are available, LSA may also help to understand efficiency, i.e. effectiveness in relation to investments. Large representative samples allow for the generalization of findings both within and across countries.
- LSA provide a data source for the study of educational contexts in general (e.g. how family, school and out-of school education interact in the development of life skills). For example, TIMSS, PIRLS, and PISA data are increasingly used by economists and social scientists to examine broader issues such as the impact of human capital on economic growth (Hanushek & Woessmann, 2009) or how to predict successful integration of migrant families (Stanat & Christensen, 2006). The database will become even more informative once these studies move into further cycles, making trend data available that cover more than a decade.

Thus, Large Scale Assessments offer three types of "products": (1) Indicators that monitor the functioning, productivity and equity of education systems. (2) Knowledge on factors that determine educational effectiveness. (3) A reliable, sustainable, comparative database that allows researchers world-wide to study basic, as well as policy-oriented, questions.

Policymakers are mainly interested in the first type of product. The policy relevance of this system monitoring enterprise is based on (a) defining and operationalizing cognitive and non-cognitive outcome measures that inform the selection and prioritisation of educational goals within participating countries, (b) examining and reporting factors that may be subject to control by policy and professional practice (so-called malleable factors) and (c) providing international benchmarks that allow policymakers to ascertain what they may learn from other countries. The selection of indicators is generally guided by policy demands. Educational policymaking must deal with the functioning of the school system (i.e. operational characteristics such as resources allocated to schools), with productivity (such as the gross level of student outcomes) and, last but not least, with equity (e.g. how resources are distributed).

Researchers, on the other hand are mainly interested in "products" (2) and (3). They tend to perceive LSA as a kind of multi-group (i.e., multi-country) educational effectiveness study. Besides describing strengths and challenges with regard to the students' performance and the conditions of teaching and schooling in participating countries, researchers – but to some extent also policymakers – intend to understand why students reach certain levels of performance. This is where context-related research comes in.

Conceptual Structure of the PISA Design for Contextual Variables

Standard models of school and teaching research conceptualize the school as a system wherein the characteristics of the context, input variables, school and instruction processes interact in "producing" student outcome. The basic structure of this Context-Input-Process-Outcome (CIPO-) model was in fact developed in the 1960s to support the design of international LSA undertaken by the IEA (Purves, 1987). Addressing the multi-level-structure of the educational systems, current versions of the framework (see Table 1) allocate input, process, and outcome characteristics at respective levels of action (i.e. system level, school level, instruction/class/teacher level, individual level).

For example, a recent version of the CIPO model, as shown in Table 1, covers practically all constructs that have been suggested for inclusion in the design of background questionnaires in the PISA 2012 study (Klieme et al., 2010). The first column displays four levels: Students, classrooms, schools and countries. The three production phases are then given in the remaining columns, i.e. inputs, processes and outcomes, respectively. As can be seen from the table, the major achievement domain in PISA 2012 will be mathematics.

The choice of constructs in LSA is based on a combination of policy priorities and research evidence. Policymakers on the PISA Governing Board decide upon the goals and research questions, while experts, building on extensive knowledge in educational effectiveness research, choose the appropriate constructs, instruments, and variables. For example, the definition of "mathematical literacy" as the most important outcome variable, and the decision to include mathematics-related attitudes and beliefs as outcome variables are both based on policy decisions, reflecting general curriculum goals and goals of the educational system shared by most participating countries. The constructs we use, however, and how these are operationalized, mainly reflect insights gained from research literature. Also, input and process variables are included if there is strong research evidence that they have an impact on the outcomes. Factors that have been demonstrated to be relevant for educational effectiveness or efficiency in the research literature are premier candidates for continuous monitoring within LSA and for incorporation into the broader system of educational indicators.

Some input factors are fairly stable and difficult to change while others can be shaped by school development activities or policy decisions. Processes are usually more malleable, at least indirectly (e.g. by teacher education and professional

Table 1 Overview of constructs covered by PISA 2012

	Input	Processes	Outcomes
Students	Gender, grade level, socio-economic status	Attendance/truancy	Mathematical literacy
	Educational career, grades	Outside-class activities – e.g. participation in after school programs	Mathematics-related attitudes, beliefs and motivation
	Immigration background	Motivation, engagement	General school-related attitudes and behaviour, e.g. commitment, truancy
	Family environment and support		
	ICT experience, attitudes, skills	Learning and thinking strategies, test taking strategies	Learning motivation, educational aspirations
	Openness, problem solving styles	Learning time (including homework and private tuition)	
Classrooms	Class size, socio-economic background and ethnic composition	Quality of instruction: structure, support, challenge	Aggregated student variables
	Teacher education/training, expertise	Opportunity to learn: implemented curriculum, assigned tasks, mathematics-related activities	
		Instructional time, grouping, assessment and feedback	
Schools	Socio-economic background and ethnic composition	Achievement orientation, shared norms, leadership, teacher morale and cooperation, professional development	Aggregated student variables
	Affluence of the community		
	School funding, public vs. private	Admission and recruitment policies, tracking, course offerings/school curriculum, evaluation	Promotion/retention and graduation rates
	School size		
	Parental involvement	Teacher-student relations, supportive environment	Attendance
Countries (systems)	Economic wealth, social (in)equality	School funding, tracking and allocation, policies for professional teacher development, support for special needs and language minority students, hiring and certification policies	Aggregated student variables
	Diversity policies	Accountability and evaluation policies, locus of decision-making	Average graduation level

From Klieme et al. (2010)

development), and outcomes reflect the effects of the inputs and processes. Note, however, that the discrimination between the three strands of variables is by no means clear-cut: Outcomes from one educational setting become input for the next, while some process aspects (e.g. learning strategies) may well be treated as input or outcome, depending on a given theoretical perspective, research design, or practical considerations.

Limitations

As Baker (2009) notes, the history of policy-making informed by international comparative studies has seen a number of short-cut conclusions, based on too simple hypotheses as to the causes of performance differences at the system level. Also, econometricians have studied a number of issues in educational productivity, but most of this work remains descriptive in nature and does not allow for causal inferences – a limit that is expressed, with regret, by Hanushek and Wössmann (2010).

For example, PISA is a yield study, assessing literacy and skills that have been accumulated over the lifespan, from early childhood through different levels of schooling until the age of 15 years. PISA does not ascertain how much learning has taken place in the secondary school where a student is presently enrolled. Such an assessment would require that the student's performance level was ascertained at the time of entering his or her present school and compared with the same student's present performance. In so doing, one would obtain a measure of progress or "value-added" in performance associated with educational experiences in the particular school. However, the PISA design does not provide any baseline measure.

The main problem with causal inferences in LSA is not a statistical or methodological one. The conditions for causal inference from quasi-experimental or survey-type data are well-known, based, e.g., on the Rubin-model of causality. Rather, the problem is substantial. The sociological theory of schooling as well as pedagogical concepts state that student achievement is the core of school education, i.e. the school expects students to strive for achievement, and its main "product" is student achievement. The process of education ("Bildung" in German) can be defined as finding an appropriate individual pathway to knowledge, competency, and expertise. Pedagogical treatments ("Erziehung" in German) need to adapt to the preconditions of learning, especially to prior achievement. Professional educators, in their daily practice, need to monitor student achievement, and change treatments accordingly. When assigning tasks, forming groups for collaborative learning, giving feedback, deciding on grade retention/promotion and other aspects of educational careers, teachers will inevitably take students' prior achievement into account. Thus, effects of these treatments cannot be estimated from cross-sectional data only, without knowing prior achievement and other factors that drive treatment assignment. Without longitudinal data, it is nearly impossible to draw causal inference in education, at least when student achievement is the dependent variable, as is the case in LSA.

Using PISA Data for Research on Educational Contexts and Their Effectiveness

Even while causal inferences are not warranted, LSA data can be put to substantial use for gaining insights in educational effectiveness: Hypotheses from general educational research can be tested in LSA, making use of broad, representative samples, high participation rates, and good measurement quality. Especially, international LSA allow for (a) studying the impact of school structure on educational processes and outcomes, and (b) checking the cross-cultural and cross-national validity of research findings.

The two papers in this part of the book are good examples for this kind of research. Both of them address research issues that are very prominent in general educational research: Angelone and Moser deal with the impact of learning time on student outcomes, which has been considered as a core variable in educational and psychological research since the seminal work by Carroll (1963). Mostafa deals with the impact of student composition on school effectiveness, which has been discussed in educational and sociological research since the Coleman Report (Coleman, 1966) came out with its surprising result on how important student background – both individual and, as subsequent re-analyses showed, compositional – is in school effectiveness. Both papers use the complex design provided by PISA to study context effects that may moderate the respective general effects. Interestingly, both refer to school types as important contextual factors: Angelone and Moser compare school types with different levels of requirements. They hypothesize that increasing learning time is less effective in advanced academic programs (rather: tracks), compared to vocational programs with lower levels of academic requirements. Mostafa assumes that “peer effects”, i.e. the impact of student composition on individual performance after controlling for individual student characteristics and organizational factors such as school funding, will be relatively strong in educational systems with strong and early stratification into different school types.

The study by Angelone and Moser is restricted to one country, and thus does not check for cross-cultural or cross-system validity of its findings. In principle, PISA would allow for testing the same hypothesis in other systems with stratified school types. Mostafa provides an example of how to do a comparison between different systems. In interpreting specifics of the five systems studied, he heavily relies on qualitative background information on structure and history of educational systems. One of the future challenges for PISA-based research on context effects is how to include system-level variables into quantitative models.

Both papers use regression models to study combined effects of predictors from different levels, such as individual, compositional, school and system level variables. This method requires careful handling of data on an aggregated level – e.g. when learning time is measured as the expected number of hours dedicated to a certain subject, i.e. on the level of the state (Canton), rather than the individual level. Also, as discussed above, one of the major problems in working with PISA data in educational effectiveness research is its cross-sectional nature. Mostafa explicitly refers to this problem when explaining that allocation to school tracks in Japan is driven by test results, and thus the “finding”, that mean socio-economic status has an effect on student performance may be an example of reverse causality. Similar arguments may hold true in the case of Swiss Cantons: The allocation of learning time to school types may be a result, rather than the cause of differences in student performance.

To sum up, both papers nicely illustrate how PISA data can be used to study general research questions, but at the same time they indicate some open questions with regard to the theoretical and methodological foundations of working with cross-sectional, comparative, multi-level data from Large Scale Assessments. Hopefully, in the future more and more educational researchers will use these data and work on advancements in educational research.

References

Baker, D. P. (2009). The invisible hand of world education culture. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 958–968). New York: Routledge.

Carroll, J. (1963). A model of school learning. *Teachers College Record, 64*, 723–733.

Coleman, J. S. (1966). *Equality of educational opportunity study*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Hanushek, E. A., & Woessmann, L. (2009). *Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation* (NBER Working Paper No. 14633). Cambridge, UK: National Bureau of Economic Research. Retrieved on March 11, 2011 from http://www.nber.org/papers/w14633.pdf

Hanushek, E. A., & Wössmann, L. (2010). *The economics of international differences in educational achievement* (NBER Working Papers 15949). Cambridge, UK: National Bureau of Economic Research. Retrieved on March 11, 2011 from http://www.nber.org/papers/w15949.pdf

Klieme, E., Backhoff, E., Blum, W., Buckley, J., Hong, Y., Kaplan, D., Levin, H., Scheerens, J., Schmidt, W., van de Vijver, F., & Vieluf, S. (2010). *Designing PISA as a sustainable database for educational policy and research: The PISA 2012 Context Questionnaire Framework*. Paris: OECD.

Purves, A. C. (1987). The evolution of the IEA: A memoir. *Comparative Education Review, 31*(1), 10–28.

Stanat, P., & Christensen, G. (2006). *Where immigrant students succeed – A comparative review of performance and engagement in PISA 2003*. Paris: Organisation of Economic Co-Operation and Economic Development.

Chapter 8
More Hours Do Not Necessarily Pay Off. The Effect of Learning Time on Student Performance at Different School Types in Switzerland

Domenico Angelone and Urs Moser

Abstract Learning time is an important determinant of academic performance. More learning time tends to be correlated with better academic performance. This article examines whether this correlation is influenced by selection in tracked schooling models. Using Swiss PISA grade 9 data for 2006, and taking science and mathematics as examples, we tested whether the correlation between learning time and performance differs among school types with different requirement levels. The results suggest that those students in school types with more advanced requirements benefit from additional learning time, but in school types with low requirements, additional learning time barely shows any positive effect.

Keywords Learning time • Performance • Selection • School type

8.1 Introduction

Learning time in teaching is an important determinant of academic performance. The positive correlation between learning time and academic performance has been empirically confirmed many times (Anderson, 1995; Fisher, 1995; Seidel & Shavelson, 2007). Evaluation of international PISA data for 2006 also shows that learning time is significant for performance in science. One additional hour of science teaching per week is associated with an increase of performance by about 9 points on the PISA-Scale (OECD, 2007, p. 263). Analyses using Germany's PISA data for 2006 lead to the same findings (Kobarg, Altmann, Wittwer, Seidel, & Prenzel, 2008; Seidel, Prenzel, Wittwer, & Schwindt, 2007).

D. Angelone, M.A. (✉) • U. Moser
Institute for Educational Evaluation, Associated Institute of the University of Zurich, Wilfriedstrasse 32, 8032 Zürich, Switzerland
e-mail: domenico.angelone@ibe.uzh.ch; urs.moser@ibe.uzh.ch

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_8,

Although it's incontestable that learning time is an important condition for the development of academic achievement, the relationship between learning time and academic achievement has not been empirically studied in connection with ability grouping, in particular with respect to school tracking at lower secondary education levels. Research on ability grouping indicates that the opportunities to learn (Baumert, Stanat, & Watermann, 2006) differ according to the attended school type. These findings suggest that the quantity of instruction has a different degree of effect depending on the school type.

Answering the question of whether or not learning time is positively correlated with academic achievement in lower level schools, and if so, to what degree, will lead to more information about the consequences of tracking on the lower secondary level, on the one hand, and to findings about the efficacy of curricular adaptations such as augmenting the quantity of instruction time in order to improve academic achievement, on the other.

The aim of the present article is to investigate whether the correlation between instruction time and subject-specific performance differs according to school type. The question is examined by testing the correlation between learning time due to curricula and performance in science and mathematics at different school types of lower secondary education in Switzerland.

With its federalist system of government, Switzerland offers a good opportunity for empirical study of these research questions. The states of Switzerland, which are called cantons, have many freedoms concerning the specific design of compulsory education. For this reason, all the cantons have different curricula—and of particular interest to us here—also different "subject tables," which outline the numbers of weekly hours of instruction by subject. The total number of compulsory hours of science instruction in the 3 years of lower secondary education varies from 228 h in the canton of Waadt to 480 h in the canton of Basel-Country. Similar differences in the number of hours of instruction also appear in the field of mathematics, which varies from 342 h in the canton of Waadt to 570 h in the canton of Freiburg (French part).

8.2 Theoretical Background and Hypotheses

In the tradition of the psychological model of education productivity, classroom learning is a function of four essential factors: students' ability and motivation, and quality and quantity of instruction (Haertel, Walberg, & Weinstein, 1983, p. 57). Fend (1998) and Helmke (2003) make a differentiation of the model, dividing between the instruction of the teacher and the use of the student (provision-utilization model). Thus, academic performance depends partly on the quantity and quality of the instruction, but also partly on how students make use of the opportunities to learn.

It is beyond controversy that the quality of instruction provided by teachers is one of the most powerful influences in learning (Hattie, 2009, p. 238). From a political perspective the quantity of instruction is relatively easy to influence by political decisions, however; opportunities to govern the quality of instruction are indirect only.

In order to explore the research questions herein, it is helpful to theoretically define the variable "quantity of instruction." First, the quantity of instruction can be defined as the intended instruction time required to teach all of the lessons in the curriculum. Second, the quantity of instruction can be defined as the implemented instruction time, referring to the number of lessons that actually took place. Third, the quantity of instruction can be defined as the maximum useable instruction time when the teacher is teaching and the student is present in the classroom. And fourth, the quantity of instruction can be defined as the time a student actively uses for learning (Helmke, 2003, p. 205).

These theoretical distinctions show that there are many aspects to be considered when interpreting the correlation between quantity of instruction respectively learning time and academic performance. Empirically, the strongest correlations with academic performance can be found considering the actively used learning time (Anderson, 1995). For the present study we define learning time as the number of lessons due to curriculum. With this definition we are just able to capture the maximal amount of time that can be used by teachers and students—but it's a variable, which can be influenced relatively simply and directly via curriculum guidelines.

Another theoretical aspect to consider when exploring the connection between learning time and academic performance is the fact that students in the lower secondary level are often divided into institutionally separated school types according to their performance level. The curricular differences between school types or classes with different requirement levels are well known, not only with respect to the quantity of instruction (e.g. Angelone & Moser, 2010), but also the quality of instruction. Students in higher-achieving classes enjoy better-quality teaching than students in lower-achieving classes (Gamoran, Nystrand, Berends, & LePore, 1995) and engage more with critical thought processes and problem-solving strategies (Oakes, 1985). In low-achieving classes, by comparison, teaching seems to be more fragmented and to proceed less quickly (Oakes, 1985; Page, 1991). Moreover, it is known from research in educational psychology that more intelligent students are, quantitatively and qualitatively, in a better position to take advantage of the offered schooling (Weinert & Hany, 2003).

Lower-achieving classes are often also particularly burdened by different factors such as a high proportion of foreign-language-speaking students, a low level of ability and performance, and a concentration of students from educationally disadvantaged families, which can lead to differential learning and development environments (Angelone, Ramseier, & Moser, 2010; Baumert & Schümer, 2002; Neumann et al., 2007). Context factors such as the social and cultural composition of the class do not directly affect learning and development environments. Mediated through value orientations of students, peers or parents, teachers' expectations and aspects of instruction methods, they can however affect opportunities to learn (Baumert et al., 2006). Students in higher-level learner groups benefit most from the differential learning and development environments. Numerous studies show that learning progress is greater in high-achieving classes than in low-achieving classes (Baumert et al., 2006; Neumann et al., 2007; Robertson & Symons, 2003).

Differences between school types—such as institutional differences in the quality of instruction and differences in the learning and development environments due to class composition—suggest that the relationship between learning time and student performance in high-achieving school types is stronger than that in low-achieving school types. Students in high-achieving school types should thus benefit more from an increase in learning time than those in low-achieving school types.

8.3 Methods

8.3.1 PISA Grade Nine Sample

To conduct the analysis we used representative data on ninth-grade students collected in the context of PISA 2006 in Switzerland. A number of cantons in Switzerland took the opportunity to collect data for PISA 2006, to complement the international sample of 15-year-olds with an additional representative national sample of students in the ninth grade. The sample of ninth graders has an advantage over the age-based international sample in that school performance can be described in dependency on characteristics of the school system.

The present analysis, therefore, is based on only ninth-grade students in cantons with a representative grade nine sample. Fifteen-year-old students who were not in grade nine were not considered in this analysis. In total, data on approximately 14,350 ninth-grade students in 14 cantons was available for the analysis. Listwise deletion of missing data reduced the analysis sample to 14,090 ninth-grade students (98% of the total ninth-grade sample).

8.3.2 Variables

8.3.2.1 Performance in Science and Mathematics

To investigate the correlation between learning time and performance, the Swiss results in science and mathematics of ninth-grade students in PISA 2006 were used. The results of the PISA test are presented on a standard scale. In 2003, the core theme of the PISA study was mathematics, whereas in 2006 science was on the focus. In PISA 2003 the scale for mathematics was standardized in such a way that the mean value for the OECD countries was 500 points with a standard deviation of 100 points (OECD, 2004). In PISA 2006 the same procedure was used for the scale for science (OECD, 2007).

8.3.2.2 Learning Time in Science and Mathematics

The variable “learning time” in the natural sciences and mathematics was collected independently from the PISA-data collection. Information on hours of learning in both subjects was taken from the curricula of the cantons (EDK, 2008). This information refers to the total learning time at the Lower Secondary Level, or, in other words, the total learning time for students in grades 7–9.

The information on learning time for mathematics could be calculated reliably because it could be derived directly from the curricula. The subjects “Geometry” and “Geometrical drawing” were counted as mathematics. The learning time in which science topics were dealt with could not be taken directly from the curricula, however. The natural sciences are not treated as one single subject (i.e. chemistry, biology etc.) and are often taught in an interdisciplinary manner; in combination with other subjects. For example, the subject area “Humans and environment” deals with more than just the core natural science disciplines. Therefore, to get corresponding information on learning time in science, experts from the cantonal departments of education first had to estimate how much time was spent on biology, chemistry, physics and geography. Due to these estimates, the instruction time in the natural sciences is approximate.

The amount of learning time in mathematics and science was collected separately for each school type (school type with advanced requirements, school type with broader requirements, and school type with basic requirements).[1] To calculate the learning time in a subject, the number of weeks of school, the number of lessons per week and the duration of the lessons were taken into account. Only the compulsory lessons in a subject were counted. The information herein relates to the 2005/2006 school year.

8.3.2.3 School Type

In Switzerland students at the lower secondary level are taught mainly in school types constituted according to performance. For the present analysis, only the results of students taught in so-called ‘type-divided’ models were taken into account. In these models, students are taught in institutionally separated school types, in accordance with their level of performance.

Fundamentally, three school types can be distinguished: the school type with basic requirements is often also referred to as ‘Realschule’, and prepares students for simple occupational training. The school type with broader requirements is called a ‘Sekundarschule’ and prepares students for more demanding occupational training or further schooling. The school type with advanced requirements is usually referred to as a ‘Gymnasium’ and prepares students for their university entry qualification (Matura). The different types of school each have their own curricula with corresponding requirements for the number of hours spent on each subject per week. The requirements due to school types and curricula also vary from canton to canton.

[1] For the school type with advanced requirements (e.g. Gymnasium), the numbers of hours given are an average of all types of “Maturität” (university entrance qualification).

Table 8.1 Descriptive statistics of dependent and independent variables

	Mean/proportion	SD	Min	Max	N
Dependent variables					
Performance in science	520	88	196	825	14,348
Performance in mathematics	542	87	216	865	14,348
Learning time in hours (grades 7–9)					
Science	329	64	228	480	14,348
Mathematics	451	51	342	570	14,348
Economic, social and cultural status					
ESCS-index	0.03	0.88	–4.36	2.77	14,293
School type					
Basic requirements	26%				14,348
Broader requirements	34%				14,348
Advanced requirements	39%				14,348
Immigration background					
Native students	78%				14,123
Second-generation students	12%				14,123
First-generation students	10%				14,123
Gender					
Male	50%				14,348

Note: PISA 2006, cantons with representative grade 9 sample. Students who cannot be allocated to a specific school type (type-divided model) were excluded from the analysis

8.3.2.4 Background Variables

In the analysis of the effect of learning time on performance, the economic, social and cultural status (ESCS), immigration background and gender of the students were statistically controlled. An ESCS index was derived from the highest occupational position of the parents, the highest educational qualification of the parents, and the possessions present in the family home (OECD, 2007, p. 333). The index shows an OECD mean of 0 and a standard deviation of 1. Students' immigration backgrounds were measured by means of a triple-level variable (OECD, 2007, p. 334): (1) native students (those students born in Switzerland or who had at least one parent born in Switzerland), (2) second-generation students (those born in Switzerland but whose parents both were born in another country) and (3) first-generation students (those born outside Switzerland and whose parents were also born in another country). The students without a migrant background constitute the reference category in the analysis. The mean values and standard deviations of all considered background variables used are presented in Table 8.1.

8.4 Statistical Methods

In order to investigate the effect of learning time on student performance we estimated OLS regressions based on weighted data (PISA final student weight). In spite of the hierarchical data structure—students are nested in schools—there were two main reasons for not using Multilevel Analysis in the present case. First there are some limitations of the use of Multilevel Models in the context of the PISA Data (OECD, 2009, pp. 221–222). Because in Switzerland schools were defined as administrative units, they can represent students from institutionally different school types that are not comparable with any restrictions. Out of the 302 schools considered in the present analysis, 34% represent students from a single school type and 64% represent students from at least two different school types. Secondly, our variable of interest, learning time, varies at the student level, since the amount of instruction time differs not only between the cantons but also according to the attended school type within a canton (cf. Table 8.1).

In order to take into account the stratified two-stage sampling design of PISA, the standard errors of the applied OLS regression coefficients were estimated using the 80 PISA replicate weights.[2] This leads to unbiased estimates of the standard errors (OECD, 2009, pp. 70–75).

8.5 Results

8.5.1 Learning Time—Comparison Between Cantons

Table 8.2 shows, for the cantons studied, the hours of learning time which students in grades 7–9 of the lower secondary level spend on science and mathematics due to curricula. The learning time differs—sometimes considerably—between cantons, but also often varies within cantons between school types.

Within school types with advanced requirements it is students in the canton of Schaffhausen who spend the most learning time on science in the lower secondary level: 477 h. In comparison, students in the same school type in the canton of Aargau spend only about half this time, 247 h, on science. The differences are similarly great between schools with broader requirements. While 480 h are used for teaching science in the canton of Basel-Country, only 240 h are spend on this in the canton of Zurich. Students in the school type with basic requirements receive 424 h of science teaching in the canton of Schaffhausen, but only 228 in the canton of Waadt. Thus, the tendency is for students in school types with advanced and broader requirements to receive more learning time for science than those in school types with basic requirements.

The differences in mathematics teaching between the cantons are not as great as those in science teaching, but are still considerable. In the school type with advanced requirements it is students in the canton of Schaffhausen who spend the most learning time in the lower secondary level, 514 h, on mathematics, and students in the canton of Waadt who spend the least amount of learning time, 342 h, on this subject.

[2] Variables W_FSTR1 to W_FSTR80, final student replicate BRR-Fay weights.

Table 8.2 Total hours of learning in science and mathematics in grades 7–9 due to curriculum

	Science			Mathematics		
	Advanced requirements	Broader requirements	Basic requirements	Advanced requirements	Broader requirements	Basic requirements
Aargau	247	463	350	463	463	556
Bern (German part)	357	304	304	410	410	410
Bern (French part)	351	351	351	410	468	468
Basel-Country	420	480	360	390	450	435
Freiburg (French-part)[a]	253	348	348	443	475	570
Geneva	318			375		
Neuenburg	293	263	263	410	439	527
St Gallen	400	383	383	467	500	500
Schaffhausen	477	424	424	514	497	497
Thurgau	375	360	360	480	510	510
Waadt	314	342	228	342	456	428
Wallis (German part)	304	253	231	459	475	475
Walis (French part)[a]	304			459		
Zurich	293	240	240	390	480	480

Note: Source: EDK (2008)—adjusted by cantonal experts. The learning time for science includes lessons in biology, chemistry, physics and geography. The learning time for mathematics also includes lessons on geometrical drawing

[a] In the French-speaking part of the canton of Wallis, and in the canton of Geneva students are not taught in a type-divided model, except for the school type with advanced requirements

In the school type with broader requirements it is students in the canton of Thurgau who spend the most time on mathematics classes: 510 h. In the German-speaking part of the canton of Bern, students of the same school type have only 410 h of mathematics classes. In the school type with basic requirements, 570 h are spent on mathematics teaching in the French-speaking part of the canton of Freiburg, while only 410 h are spent on it in the German-speaking part of the canton of Bern. In contrast to teaching in science, the tendency in mathematics is to offer students with basic requirements more learning time.

The 14 cantons thus present us with a relatively homogeneous sample—especially as compared to a country analysis, where the different framework conditions, such as quality of the education system, take on much greater importance—but a sample in which the instruction times for science and mathematics vary.

8.5.2 Effects of Learning Time on Performance in Science and Mathematics

In order to investigate the correlation between learning time and student performance at different school types, OLS regressions were estimated. The dependent variables were the performance levels of the students in science and mathematics at the end of grade 9. The hours of learning time available to students from grades 7 to 9 were incorporated as the predictor. By further taking into account the square of the hours of learning time as a predictor, it becomes possible to estimate non-linear correlations between learning time and performance. In order to test the correlation between learning time and performance dependent on school type; interaction terms between the school type and the learning time were established and introduced into the models. The reference category was constituted in each case by the school type with advanced requirements (Gymnasium). Alongside the ESCS index, the students' immigration background, gender and interaction terms between these variables anwd the school type were included in the analysis as control variables.

In the following discussion of the results, it should be noted that the cross-sectional design of PISA and regression analysis do not allow for causal inferences. The present analysis allows simply for the detection of observable correlations between instruction time due to curricula and student performance at different school types, adjusted for school type and background variables.

The results of the regression analysis shown in Table 8.3 suggest that learning time has a different effect on students' performance depending on the subject and the school type. For the school type with advanced requirements—our reference category—learning time is positively correlated with students' performance in science. This correlation is charted in the left-hand section of Fig. 8.1. This suggests that, from a starting level of around 300 h of learning time (vertex of the curve), more hours clearly have a positive effect on student performance. When learning time at the lower secondary level is increased from 300 to 400 h, student performance

Table 8.3 Results of the regression analysis on the effect of learning time on performance in science and mathematics

	Model Science		Model Mathematics	
Effects of learning time	*B*	*SE*	*B*	*SE*
Learning time (in hours per year)	0.156***	0.034	0.397***	0.077
Learning time (squared)	0.003***	0.000	0.001	0.001
Learning time * Broader requirements	–0.131*	0.050	–0.271*	0.108
Learning time (squared) * Broader requirements	–0.003***	0.001	0.015***	0.002
Learning time * Basic requirements	–0.092	0.055	–0.281*	0.120
Learning time (squared) * Basic requirements	–0.004***	0.001	–0.003*	0.001
Broader requirements	–50.705***	3.442	–85.743***	5.066
Basic requirements	–114.956***	4.866	–133.969***	5.555
Effects of background variables				
Male	15.286***	2.525	23.213***	2.606
Male * Broader requirements	9.724*	3.838	7.249	3.880
Male * Basic requirements	9.511*	3.806	10.188**	3.510
First-Generation	–30.986**	5.912	–27.31***	5.490
First-Generation * Broader requirements	–21.185*	8.377	–15.690	8.480
First Generation * Basic requirements	–22.778**	6.806	–19.732**	7.090
Second-Generation	–42.653***	4.847	–32.485***	4.728
Second-Generation * Broader requirements	–2.261	6.401	–1.137	6.321
Second Generation & Basic requirements	–0.722	6.509	–1.076	7.045
ESCS-Index	10.892***	1.947	9.822***	1.935
ESCS-Index * Broader requirements	–5.319*	2.430	–5.505*	2.578
ESCS-Index * Basic requirements	–1.947	3.363	–4.476	3.405
Constant	573.878***	3.064	603.670***	2.973
R^2	.454		.448	

Note: N = 14,090. Linear regression models based on weighted data, calculation with 5 "plausible values" and estimation of standard errors by use of the 80 PISA replicate weights (OECD, 2009). The variables "Learning time" and "ESCS-Index" are centered around the overall mean

* p<.05, ** p<.01, *** p<.001

wimproves by 32 points. The curvilinear form of the correlation graph shows that more hours have a stronger positive effect on performance if the learning time starts at a high level than if it starts at a low level. Figure 8.1 also suggests, however, that at a very low level a slight increase in science learning time does not bring about any improvement in performance.

For the school type with broader requirements, however, no statistically significant correlation between learning time in science and student performance can be

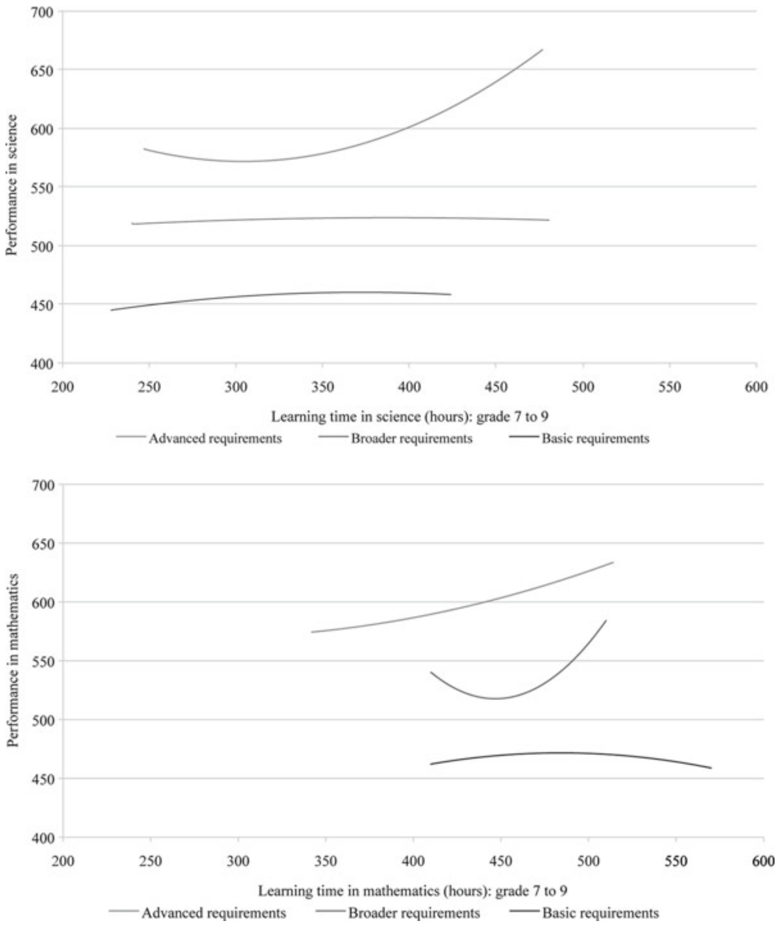


Fig. 8.1 Effect of learning time on performance in science and mathematics Note. Estimated effects of learning time on performance on the basis of the regression analysis in Table 8.3. The effects are presented for the observed range of learning time values. The expected performance is shown as an example for students without a migrant background and with average economic, social and cultural status (ESCS-Index)

detected.[3] In contrast to the school type with advanced requirements, more hours do not pay off in this school type. In the school type with broader requirements, as Fig. 8.1 shows, performance in science is expected to be around 525 points, regardless of the available learning time.

[3] Wald test is used to test the joint significance for the linear and quadratic terms of learning time for the school type with broader requirements: B(Learning time) + B(Learning time * Broader requirements) = (Learning time squared) + B(Learning time squared * Broader requirements) = 0, χ^2 (2) = .49, Prob > χ^2 = .783.

For the school type with basic requirements, there is some evidence of a statistically significant correlation between learning time and performance in science.[4] However, the effect of learning time is relatively weak. When learning time is increased by 100 h, student performance improves by only 10 points.[5]

The findings in science suggest that more learning time is not necessarily associated with better performance. In this analysis, a positive and significant correlation between learning time and student performance can only be found for the school type with advanced requirements. For the two school types with lower performance requirements, more learning time has no effect or only a limited effect on student performance. The R^2 indicates that the fitted model explains 45.4% of the variation in science performance. After accounting for school type and background variables, however; the inclusion of learning time improves the model by just 1%. This shows that the partial explanatory contribution of learning time is of minor importance. Let us now consider the results for mathematics.

For the school type with advanced requirements, there is evidence that learning time has a positive and significant effect on student performance (cf. Table 8.3). The effect is charted in the right-hand section of Fig. 8.1. When learning time is increased by 100 h, performance in mathematics can be expected to improve by around 30 points. As Fig. 8.1 suggests, performance in mathematics increases in an almost linear relation to increased learning time. The quadratic term of learning time is not statistically significant in the school type with advanced requirements.

For the school type with broader requirements, more hours of mathematics teaching clearly pay off.[6] Figure 8.1 shows that, starting from a learning time of 450 h (vertex of the curve), additional hours are accompanied by a marked improvement in mathematics performance. When learning time is increased from 450 to 510 h, the highest learning time observed in the sample, mathematical performance improves by 66 points. The results, however, also indicate that student performance depends on other factors as well as learning time, factors which were not able to be taken into consideration in the present analysis. This is shown by results from the students of one particular canton (the German-speaking part of the canton of Bern), who perform relatively well in mathematics despite having the least learning time at the lower secondary level (410 h).

[4] Wald test is used to test the joint significance for the linear and quadratic terms of learning time for the school type with basic requirements: B(Learning time) + B(Learning time * Basic requirements) = B(Learning time squared) + B(Learning time squared * Basic requirements) = 0, $\chi^2 = 7.24$, Prob > $\chi^2 = .027$.

[5] In the school type with basic requirements the correlation between learning time and science performance is linear.

[6] Wald test is used to test the joint significance for the linear and quadratic terms of learning time for the school type with broader requirements: B(Learning time) + B(Learning time * Broader requirements) = B(Learning time squared) + B(Learning time squared * Broader requirements) = 0, $\chi^2 = 153.05$, Prob > chi2 = .000.

For the school type with basic requirements, finally, the results of the regression analysis show no statistically significant correlation between learning time and mathematics performance.[7] As illustrated in Fig. 8.1, mathematics performance changes only negligibly as learning time increases. In this school type, more hours of teaching do not pay off.

In mathematics, the findings on the effect of learning time on student performance are more clear-cut than in science. With the exception of the school type with basic requirements, there is evidence of positive correlations between the available learning time and student performance. The fitted model explains 45.4% of the variation in mathematic performance. After accounting for school type and background variables, the inclusion of the learning time improves the model by 3%, which is slightly more than in the science model (1%).

8.6 Conclusion

The results of this analysis suggest that more learning time leads to better student performance, and confirm the significance of learning time for academic performance (Anderson, 1995; Fisher, 1995; Seidel & Shavelson, 2007; Seidel et al., 2007). The effects of learning time vary, however, according to subject and school type.

In science, learning time can—on the basis of our analysis—only be proven to have a positive and significant correlation with student performance in the school type with advanced requirements. When learning time in science is increased by 100 h at the lower secondary level—with 40 weeks of school in a year, this corresponds to one additional 50-min lesson per week—student performance improves by about 30 points.

In mathematics, the results are more clear-cut than in science. More learning time is correlated with better mathematics performance, both in the school type with advanced requirements and the school type with broader requirements. In the school type with advanced requirements the correlation between learning time and mathematics performance is comparable in strength to that in science. In the school type with broader requirements, more learning time in mathematics pays off markedly more than in science. If we exclude one canton (the German part of the canton of Bern), mathematics performance improves by 66 points with a 60-h increase in learning time.

One explanation for the closer connection between learning time and mathematics performance could lie in the fact that learning time can be more reliably measured on the basis of the curriculum in mathematics teaching than in science. Science is,

[7] Wald test is used to test the joint significance for the linear and quadratic terms of learning time for the school type with basic requirements: B(Learning time) + B(Learning time * Basic requirements) = B(Learning time squared) + B(Learning time squared * Basic requirements) = 0, $\chi^2 = 2.20$, Prob > $\chi^2 = .333$.

in part, taught in an interdisciplinary manner. Moreover, the guidelines for content are more precise in mathematics than in science.

More learning time does not always pay off, however. In science, little or no effects of learning time on student performance can be detected for the school type with basic requirements or that with broader requirements. Nor is there evidence, for the school type with basic requirements, that learning time affects student performance in mathematics.

This result suggests that conditions for learning and development are less optimal in school types with basic requirements than in school types with advanced requirements. More learning time does not seem to be of greater importance for students in school types with basic requirements. At the same time, the content with regard to the quality of teaching has to be improved. More of the same can be the right or the wrong strategy to improve achievement. Increasing the amount of lessons is not an effective strategy to improve performance of all students without considering the quality of learning and contextual factors, such as the social and cultural composition of classes.

References

Anderson, L. W. (1995). Time: Allocated and instructional. In L. W. Anderson (Ed.), *International encyclopedia of teaching and teacher education* (pp. 204–207). Oxford: Pergamon.

Angelone, D., & Moser, U. (2010). Unterrichtszeit, Unterrichtsorganisation und Kompetenzen. In D. Angelone, E. Ramseier, C. Brühwiler, V. Morger, U. Moser, & E. Steiner (Eds.), *PISA 2006 in der Schweiz. Die Kompetenzen der Schülerinnen und Schüler im kantonalen Vergleich* (pp. 100–117). Oberentfelden: Sauerländer Verlage AG.

Angelone, D., Ramseier, E., & Moser, U. (2010). Schulstruktur und Selektivität. In D. Angelone, E. Ramseier, C. Brühwiler, V. Morger, U. Moser, & E. Steiner (Eds.), *PISA 2006 in der Schweiz. Die Kompetenzen der Schülerinnen und Schüler im kantonalen Vergleich* (pp. 72–99). Oberentfelden: Sauerländer Verlage AG.

Baumert, J., & Schümer, G. (2002). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb im nationalen Vergleich. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, et al. (Eds.), *PISA 2000: Die Länder der Bundesrepublik Deutschland im Vergleich* (pp. 159–201). Opladen: Leske + Budrich.

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen. Vertiefende Analysen im Rahmen von PISA 2000* (pp. 95–188). Wiesbaden: Verlag für Sozialwissenschaften.

EDK. (2008). *Stundentafeln in der Volkschule*. Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren (EDK).

Fend, H. (1998). *Qualität im Bildungswesen. Schulforschung zu Systembedingungen, Schulprofilen und Lehrerleistung*. Weinheim: Juventa.

Fisher, C. W. (1995). Academic learning time. In L. W. Anderson (Ed.), *International encyclopedia of teaching and teacher education* (pp. 430–434). Oxford: Pergamon.

Gamoran, A., Nystrand, M., Berends, M., & LePore, P. C. (1995). An organizational analysis of the effects of ability grouping. *American Educational Research Journal, 32*(4), 687–715.

Haertel, G. D., Walberg, H. J., & Weinstein, T. (1983). Psychological models of educational performance: A theoretical synthesis of constructs. *Review of Educational Research, 53*(1), 75–91.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Helmke, A. (2003). *Unterrichtsqualität—erfassen, bewerten, verbessern*. Seelze: Kallmeyersche Verlagsbuchhandlung.

Kobarg, M., Altmann, U., Wittwer, J., Seidel, T., & Prenzel, M. (2008). Naturwissenschaftlicher Unterricht im Ländervergleich. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, et al. (Eds.), *PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich* (pp. 265–296). Münster/New York/München u.a: Waxmann.

Neumann, M., Schnyder, I., Trautwein, U., Niggli, A., Lüdtke, O., & Cathomas, R. (2007). Schulformen als differenzielle Lernmilieus: Institutionelle und kompositionelle Effekte auf die Leistungsentwicklung im Fach Französisch. *Zeitschrift für Erziehungswissenschaft, 10*(3), 399–420.

Oakes, J. (1985). *Keeping track: How schools structure inequality*. New Haven, CT: Yale University Press.

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

OECD. (2007). *PISA 2006. Science competencies for tomorrow's world: Vol. 1: Analysis*. Paris: OECD.

OECD. (2009). *PISA 2006. Data analysis manual. SPSS* (2nd ed.). Paris: OECD.

Page, R. N. (1991). *Lower track classrooms: A curricular and cultural perspective*. New York: Teachers College Press.

Robertson, D., & Symons, J. (2003). Do peer groups matter? peer group versus schooling effects on academic attainment. *Economica, 70*(277), 31–53.

Seidel, T., Prenzel, M., Wittwer, J., & Schwindt, K. (2007). Unterricht in den Naturwissenschaften. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme, et al. (Eds.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* (pp. 147–180). Münster: Waxmann.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499.

Weinert, F. E., & Hany, E. A. (2003). The stability of individual differences in intellectual development: Empirical evidence, theoretical problems, and New research questions. In R. J. Sternberg & J. Lautrey (Eds.), *Models of intelligence: International perspectives* (pp. 169–181). Washington, DC: American Psychological Association.

Chapter 9
The Anatomy of Inequalities in Educational Achievements: An International Investigation Using PISA Data

Tarek Mostafa

Abstract This chapter analyses the mechanisms of stratification and inequalities in educational achievements. The objective is to determine how stratification leads to unequal educational outcomes and how inequalities are channeled through student characteristics, school characteristics and peer effects. This analysis is undertaken in five countries differentiated by their schooling systems. The countries are Japan, the UK, Italy, Germany and Finland, and the dataset used is PISA 2003. The analysis consists of a multilevel econometric model used to explain variations in performance scores. The explanatory variables are student, school and peer characteristics. The institutional context of each education system is used to interpret the results and to describe how inequalities arise. In the last section, policy implications, based on the regression results, are derived.

Keywords Educational stratification • Achievement inequalities • Comparative analysis of education systems • Multilevel modelling

9.1 Introduction

Reducing inequalities in educational attainments has become a major preoccupation of educational reforms. Recent studies—especially the OECD's "Education at a Glance"—proved the existence of large disparities in outcomes and subsequently triggered a heightened interest in policy evaluation and international comparisons.

T. Mostafa (✉)
Centre for Learning and Life Chances in Knowledge Economies
and Societies (LLAKES), Institute of Education, University of London,
20 Bedford Way, London WC1H 0AL, England, UK
e-mail: T.Mostafa@ioe.ac.uk

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_9,

According to the traditional approach, the level of inequality is defined as the strength of the impact of social background on educational attainments. This definition is implicit in some of the empirical literature and in the international comparative reports on education, such as the "Education at a Glance" and the "PISA Reports, OECD 2003" (OECD, 2003a, 2003b, 2003c). Nonetheless, reality is more complicated. Educational achievements are not the simple direct product of social backgrounds, and in general the latter operates indirectly through intricate stratification mechanisms. For instance, students whose parents are highly educated have a tendency to perform better at school. Similarly, the same students are more likely to be attending better quality schools. Hence, performance scores can be the direct outcome of particular social characteristics (e.g. parental education) or the indirect outcome that transits through school choice. It should be noted, that stratification means that students of similar type are shepherded into the same schools. In this case, students from advantaged households will socialize with students from the same group. Further, peer quality usually coincides with other favorable school characteristics such as better schooling climate, better teacher quality and instruction techniques. In conclusion, inequalities should no longer be considered as the mere impact of students' social background on their achievements, since stratification-determined school characteristics are likely to be a source of inequality too.

Moreover, the strength of stratification is not the same across countries and therefore its impact on achievements may vary according to the institutional context of each education system (e.g. comprehensiveness vs. early selection). As a consequence, the empirical analysis must consider several countries known for their contextual differences. The objective of this chapter is to study thoroughly the mechanisms of inequality in attainments by assessing the direct effects of household characteristics and the indirect effects resulting from student sorting between schools.

It should be noted that the theoretical literature on stratification is recent and dates back to the early 1970s with the founding articles of Barzel (1973) and Stiglitz (1974). The major developments occurred in the 1990s, when two distinct bodies of literature emerged. The first studied spatial stratification between jurisdictions and neighborhoods. It includes Westhoff (1977), Rose-Ackerman (1979), De Bartolome (1990), Epple, Filimon, and Romer (1993), Nechyba (1997), Epple and Platt (1998) and Fernandez and Rogerson (1996). The second studied educational stratification between public and private schools. It includes Arnott and Rowse (1987), Epple and Romano (1998, 2006) and Nechyba (2003). The empirical literature includes a variety of studies that assess the determinants of achievements, such as peer effects, students' ethnicity and immigrant status, students' socioeconomic backgrounds, and school and teacher characteristics. Hanushek and Welch (2006) provide a good coverage of the studies of interest.

The chapter is organized as follows: In the first section, the PISA 2003 dataset, the countries and the chosen variables are presented. In the second, the econometric model is discussed. In the third, the regression results are interpreted. And finally in the last, policy implications are derived.

9.2 Data, Countries and Variables

9.2.1 Data

In this chapter, the OECD Programme for International Student Assessment dataset is used. The major advantages of using it are the following. Firstly, the dataset is very convenient for international comparisons, since a large number of countries with different education systems are included. Secondly, a wide array of student and school characteristics are accounted for. Thirdly, the major subject of assessment in PISA 2003 is mathematics which is more universal than reading because it is not culturally specific or subject to cultural relativity. Fourthly, PISA uses an innovative concept of literacy which stresses the importance of certain skills for adult life instead of assessing the mastery of a particular curriculum. Fifthly, assessed students are aged between 15 years and 3 months and 16 years and 2 months, regardless of the grade in which they are enrolled. This coverage helps measuring the extent to which knowledge is acquired independently of the structure of national school systems (e.g. entry ages, grade repetition rules, etc.). In addition to this, the structure of the PISA data allows the use of sophisticated statistical methods such as multilevel models. It should be noted that before undertaking any analysis, the dataset was imputed using multiple imputations with a Marcov Chain Monte Carlo procedure in order to make it more efficient for econometric analyses. For a complete description of the MCMC method see Gill (2008) and Robert and Casella (2004).

9.2.2 Countries

Five countries with different schooling systems were selected. These are: Germany representing German speaking countries (known for early selection), Italy representing the Mediterranean countries (Italy is known for its selection at the end of lower secondary schooling and for high geographical disparities), Finland representing the Nordic countries (known for their comprehensiveness), the UK for the English speaking ones (known for the liberal organization of education) and finally Japan for East Asia (Japan is known for its strong selection at the end of the lower secondary phase). This selection is motivated by two arguments. First, it is more reasonable to select few countries representative of major schooling systems than to select all countries with some being irrelevant to the objective of the analysis. Second, it is impractical to work with the entire sample of countries because of the lack of space and the thoroughness of the analyses to be conducted. The selection is based on the Green, Preston, and Janmaat (2006) typology of education systems. Mostafa (2009) provides a thorough description of the five education systems backed by descriptive statistics.

One should keep in mind that the sampled Japanese and Italian students have already finished the lower secondary phase and have been stratified into upper secondary schools. In particular, Japanese students are sorted according to placement tests administered by the prefectural boards of education and according to their previous records. This is probably the reason why the between-school differences are high in both countries.

9.2.3 Variables

The variables used in the regression analyses are grouped in three categories. They account for multiple dimensions such as: students' socio-economic backgrounds, student motivation and interest, school funding, school environment and peer effects. They are:

9.2.4 Student Characteristics

ESCS: Economic, social and cultural status of the household.
COMPHOME: An indicator on computer facilities at home.
INTMAT: An indicator on interest in mathematics.
ANXMAT: An indicator on anxiety in mathematics.
DISCLIM: An indicator on the perception of discipline in a school.
ETR: A dummy variable taking the value of one if a student is a first generation student or a non-native. Henceforth, this category is simply called "non-natives". Note that ETR is not a measure of ethnic belonging.
Grade: a variable that controls for the grade in which a student is enrolled. Since PISA is age based and since all students were evaluated using the same test items, students' grades have to be considered, in order to control for their effect on achievements.

9.2.5 Peer Effects, School Aggregates of Individual Characteristics

DESCS: School average ESCS, depicting economic, social and cultural peer effects.
VARESCS: The within-school dispersion of ESCS, reflecting nonlinearities in peer effects (the impact of social and economic diversity).
DCOMPH: School average COMPHOME, depicting the possession of computer facilities peer effects.
DINTMAT: School average INTMAT, depicting peer effects resulting from a generalized interest and enjoyment of mathematics within a school.

DANXMAT: School average ANXMAT, depicting peer effects resulting from a generalized feeling of anxiety and helplessness in mathematics.
DDISCL: School average DISCLIM, depicting the impact of a generalized perception of discipline in a school.
DETR: The percentage of non-natives or first generation students in a school.

9.2.6 Pure School Characteristics

Compweb: The proportion of computers connected to the web in a school.
Mactiv: The number of activities used to promote engagement with mathematics in a school.
Mstrel: An index measuring poor student teacher relations.
Tcshort: An index measuring principals' perception of potential factors hindering the recruitment of new teachers, and hence instruction.
Tcmorale: An index depicting principals' perception of teacher morale and commitment.
Teacbeha: An index depicting principals' perception of teacher-related factors hindering instruction or negatively affecting school climate.
Private: A dummy variable taking the value of one if a school is private (private dependent and independent schools are combined into this variable). Note that each of the selected countries, in fact, has only one of the two types of private schools. Thus, the two types have to be combined since estimation is not possible if the frequency of one of the types is close to zero. However, the interpretation of the results is made according to the predominant type.
Scmatedu: The quality of educational infrastructure in a school as perceived by the principal.
Academic: A dummy variable taking the value of one if a school selects its students according to their academic records.

9.3 Multilevel Modeling

The model to be estimated is the following:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \gamma_1 \overline{X}_{\bullet j} + \gamma_2 K_j + \varepsilon_{ij}$$

with

$$\beta_{0j} = c + V_j \text{ and } \beta_{1j} = \beta + \mu_j$$

X_{ij} is a vector of student characteristics (student *i* attending school *j*), $\overline{X}_{\bullet j}$ is a vector of peer effects (school aggregates of student characteristics), and K_j is a vector of pure school characteristics (e.g. funding, school environment, etc.). ε_{ij} are the residuals of the model, they follow a normal distribution, with zero mean and a constant variance of σ^2, $\varepsilon_{ij} \sim \mathrm{N}(0,\sigma^2)$. When the intercept and the regression coefficient on X_{ij} are replaced by their values, the equation becomes:

$$Y_{ij} = c + \beta X_{ij} + \lambda_1 \overline{X}_{\bullet j} + \gamma_2 K_j + V_j + \mu_j X_{ij} + \varepsilon_{ij}$$

with $\beta\gamma_1\gamma_2$ being the regression coefficients on student, per, and school characteristics respectively.

Note that, the intercept is divided into two elements: c is the overall intercept, which is constant for all schools and equal to the average of the intercepts β_{0j}, and a random part V_j, denoting school *j* departure from the overall intercept, which can also be seen as a unique effect of school *j* on the average intercept (Raudenbush & Bryk, 2002). V_j can be considered as comprising the unobserved school characteristics. V_j is assumed to have a zero mean and a variance of τ_0^2. $V_j \sim \mathrm{N}(0,\tau_0^2)$. Similarly, the slope on student variables is divided into two elements: β is the overall regression coefficient, equal to the average of regression coefficients β_{1j}, and a random part μ_j, denoting school j departure from the overall regression coefficient, which can also be seen as a unique effect of school j on the slope of *X* (Raudenbush & Bryk, 2002). μ_j is assumed to have mean of zero and a variance of $\tau_1^2 . \mu_j \sim \mathrm{N}(0,\tau_1^2)$. Notice that V_j and μ_j are treated as random errors following normal distributions. The variances on V_j and μ_j are called between school variances. This model must satisfy a number of independence and normality properties; these are enumerated in Mostafa (2009).

Note that in Mostafa (2009), endogeneity problems were assessed by applying the Hausman test on several variants of the aforementioned model. Furthermore, homoscedasticity and the independence of the error terms were also assessed using residual scatter plots and Q-Q plots. The major finding is that, when peer effects are omitted, the model does not pass the Hausman test in the five selected countries. This confirmed that peer effects are a major product of stratification and that their omission leads to correlations between the error term and the included student characteristics and, hence, biased results. The most reliable model is the one that controls for the following three vectors: student characteristics, peer effects, and school variables. In what follows, only the results from the aforementioned general model are interpreted, since it passed the Hausman, the homoscedasticity, and the independence tests. Moreover, the sensitivity of the model with imputed data was tested against several regressions estimated without imputations, and with different imputation methods. In all cases, the various regressions generated results of similar magnitude and statistical significance, confirming that they are not driven by imputation techniques. Note that the multilevel model is estimated using a maximum likelihood procedure.

9.4 Results

Before interpreting the results, it is useful to start with a statistical definition of inequalities. Inequalities do not exist when—in a regression analysis—all student and school variables have insignificant effects. In other words, the variables that

Table 9.1 The variance components

	Germany	Finland	UK	Italy	Japan
Total variance	3659.08	3875.95	4187.87	3950.43	4431.33
Within variance	2436.41	3647.92	3771.91	2771.36	3661.98
Between variance	1222.67	228.02	415.96	1179.06	769.35
%Between/Total	33.41	5.88	9.93	29.85	17.36

may explain differences in performance scores are still hiding in the unobserved component (the error term). These could be student competencies (e.g., IQ). This situation is a perfect meritocracy, where the surrounding environment of a student—whether at home or at school—does not affect his achievements. Of course, this situation does not exist, but it is useful to consider it as a benchmark against which countries are compared. In what follows, the results on key variables are interpreted (Tables 9.1 and 9.2).

Before interpreting the results, I should note that the model for Japan was estimated without three variables: Grade, ETR, and Detr. This is done because the frequencies for these variables are very small (Japan has a very limited immigrant population and almost all students are in the first grade of high school), and the variables had a completely insignificant effect on performance scores. This omission had no effect on the magnitude and significance of other regression coefficients.

9.4.1 Social Status and Social Peer Effects

ESCS is the most important dimension according to which stratification operates. In most countries, it is at the centre of educational policies, since one of the objectives is to ensure equality of opportunity in the access to education. ESCS is statistically significant across all countries except for Japan. Finland has the highest value on the regression coefficient, with an increase of 25 points in performance scores caused by an increase of one unit of ESCS. Finland is followed by the UK, Germany, and Italy. The coefficients on school average ESCS are all statistically significant at the level of 1% for all countries except for Finland which has an insignificant result. The highest value on the coefficient is for Japan, followed by Germany, The UK, and Italy.

The results for Finland seem to be counterintuitive for a comprehensive schooling system. However, when average ESCS (social peer effects) is taken into account, a full picture will emerge. Comprehensiveness in Finland is associated with high levels of homogeneity between schools. Therefore, the impact of school variables, including peer effects on performance scores, is expected to be small. As a consequence, the only factors that would explain the variation in performance scores are student characteristics, such as ESCS. Furthermore, this high value on the regression coefficient is not alarming since Finland has the lowest national dispersion of ESCS. In other words, even if the slope is important, there are limited variations on ESCS to cause high inequalities in performance.

Table 9.2 The regression results

	Germany			Finland			UK			Italy			Japan		
Variables	Coef.		SE	Coef.		SE	Coef.		SE	Coef.		SE	Coef.		SE
Intercept	120.2		15.4	132.8		25.1	302.0		18.9	181.5		16.3	600.8		31.8
Grade	36.4	***	1.3	43.3	***	2.5	16.0	***	1.5	34.6	***	1.3			
ESCS	8.2	***	1.1	25.0	***	1.3	21.8	***	1.2	1.6	**	0.9	1.8		1.7
COMPHOME	0.5		1.1	−2.6	**	1.2	7.8	***	1.2	8.4	***	0.8	3.1	**	1.3
INTMAT	5.3	***	0.9	15.0	***	1.2	−1.1		1.2	6.5	***	0.8	9.8	***	1.2
ANXMAT	−16.8	***	0.9	−31.4	***	1.1	−25.1	***	1.2	−21.7	***	0.9	−8.5	***	1.2
DISCLIM	0.9		0.8	0.9		1.1	12.4	***	0.9	1.7	**	0.8	2.6	**	1.3
ETR	−21.8	***	2.9	−45.0	***	8.0	−2.2		4.3	5.1		5.5			
DESCS	47.9	***	9.3	6.3		6.1	46.9	***	6.5	31.5	***	6.1	91.9	***	14.4
VARESCS	1.4		9.5	−0.4		9.4	22.1	***	7.1	−33.6	***	8.2	43.2	**	20.2
Dcomph	41.6	***	14.6	−3.7		9.1	−2.4		9.2	50.0	***	8.5	17.0		14.9
Dintmat	−23.8	**	10.5	−11.1		8.5	−22.4	***	7.2	−17.1	***	6.3	45.9	***	12.6
Danxmat	−22.5	**	11.6	4.5		9.2	−14.4	*	7.6	−51.2	***	9.2	38.6	***	13.2
Ddiscl	27.1	***	7.6	−1.8		5.1	13.0	***	4.7	14.7	***	5.2	29.9	***	7.0
Detr	25.0		21.3	51.4		51.4	−42.7	***	15.1	72.1	***	24.2			
COMPWEB	9.8		8.1	13.4		10.4	18.2	***	6.9	6.7		5.7	−1.9		9.2
MACTIV	6.8	*	4.0	1.9		2.7	−1.9		1.6	7.8	***	2.3	1.0		4.6
MSTREL	44.2		54.3	−111.7	**	56.4	−103.2	**	43.4	−83.0	*	46.7	−108	**	51.4
TCSHORT	−5.2	*	3.7	0.01		2.4	−1.3		1.8	3.2	*	2.5	−0.4		3.1
TCMORALE	6.5	*	3.4	2.6	*	2.1	−0.8		1.8	−1.5		2.4	−2.1		3.3
TEACBEHA	−11.3	**	4.8	−2.5		2.3	0.1		2.4	1.1		2.3	7.8		4.0
private	−2.5		11.3	−20.3	***	6.8	13.2	**	5.2	−38.3	***	8.9	−38.3	***	6.8
SCMATEDU	6.2	*	3.4	−0.5		2.3	1.2		1.8	5.5	***	2.2	−5.1		3.2
Academic	5.0		5.9	11.0	**	5.1	15.8	***	4.3	−5.4		4.1	−47.9		25.3

***, **, and * stand for significance at the level of 1%, 5%, and 10% respectively

In Germany, the high levels of stratification and social inequalities in the access to education are translated into inequalities in performance scores. The ESCS of a student determines the school in which he is enrolled as well as a certain proportion of his performance. Hence, a student with a low level of ESCS is likely to be streamed into "Hauptschulen" where other students with similar levels of ESCS are enrolled. Since ESCS has an important effect on performance, low ESCS students will get lower results. And since average ESCS in a school also has an important effect on performance, low ESCS students enrolled in schools with low average ESCS are likely to have lower performances. In Germany, the selective school system is a generator of inequalities, since it allows ESCS to play fully through its direct household effect and through its indirect school effect. However, the German system cannot be understood unless the labor market is considered. Germany retains a strong apprenticeship system through which low ESCS students are shepherded into vocational tracks and educational inequalities are absorbed by the labor force. In addition to this, attending professional schools is not regarded as a sign of failure and is not associated with a socially negative stature.

The UK as well has an important effect of ESCS and DESCS on student performance scores. This is perhaps the result of the unachieved comprehensivization of the British education system. In fact, both student level ESCS and school level DESCS have significant and important effects. The UK resembles to Germany on this aspect, even though it does not have a strong apprenticeship system.

Japan also has high levels of social stratification. However, inequalities operate differently than in Germany. On the one hand, a student's ESCS has an economically and statistically insignificant direct effect on his performance; while on the other hand, school average ESCS has a very important and significant effect on performance scores. These results reflect the role of schools in the Japanese education system. Schools assume multiple roles; they are the place for the acquisition of knowledge and for the socialization of children. Instruction is organized in a way to maximize peer effects and to intensify the interactions between students. Hence, peer effects are expected to be important (see Green, 1997, 1999). However, this result should not be used to establish a complete causality going from DESCS towards performance scores. Since the sampled Japanese students have been together for only 3 months after being tracked into different high schools, the notion of peer effects in the case of Japan should be explained carefully. In fact, school average ESCS is the result of stratification according to achievements on the placement test undertaken at the end of the lower secondary phase. Thus, it is reasonable to acknowledge that the causality between DESCS and performance scores in the case of Japan might work in both directions. In other words, DESCS can be seen as peer effects affecting performance scores on the standardized PISA test as well as the result of performance scores on the placement test undertaken 3 months earlier.

Italy has very similar results to Japan, even if they are quantitatively lower. It can also be described as a country with a high level of stratification, where the social status of a school determines the performance of students. A student's own ESCS has a very low effect on his performance, while the status of the school is much more important. The same interpretations made for Japan apply for Italy, except that the impact of social peer effects is significantly lower. It should be noted, that all 15 year-old PISA students in Italy and Japan are in the upper secondary phase, which is differentiated and not comprehensive. This may help explain the high between school dispersions and the high significance of the coefficients on school variables.

Another interesting finding is that peer effects are non-linear in their means in three of the five countries (VARESCS is significant in the UK, Italy and Japan). This finding confirms my theoretical assumptions and determines how performance scores react to changes in social diversity. In the UK and Japan, an increase in the within-school dispersion of ESCS enhances performance while the reverse is true in Italy.

Note that when the results in this analysis are compared to those published in the PISA 2009 report (Volume II, p.195), it is possible to see that they are of similar magnitude and significance, even though those in the PISA report are slightly higher for all countries. This is due to the fact that in the report, only student ESCS and average school ESCS were controlled for. However in this analysis a wide array of controls were included. Despite that, my results are similar and they validate those found in PISA 2009.

9.4.2 Funding and School Characteristics

Several proxies of school funding were retained: the proportion of computers connected to the web, the number of activities promoting mathematics, teacher shortages, and the quality of educational infrastructures. The level of significance and the value on the coefficients vary between countries. For instance, in Germany, the number of activities promoting mathematics, teacher shortages, and the quality of educational infrastructures have significant effects; while in Finland and Japan none of the coefficients is significant. In the UK, only COMPWEB has a significant effect, while in Italy, the number of activities promoting mathematics, teacher shortages, and the quality of educational infrastructures have significant effects. All the coefficients are of the expected sign. These results reflect the high between-school disparities in Germany and Italy and the homogeneity of schools in Finland. When it comes to school environment, only MSTREL (poor teacher student relations) has a significant and negative effect across all countries. Teacher morale has a positive and significant effect in Germany and Finland, while negative teacher behaviour has a negative and significant effect in Germany.

9.4.3 Private Schooling

Private education is also an important determinant of performance. It has a significant effect across all countries except in Germany. At the first sight, the results seem to be counter intuitive since the sign on the coefficient is negative except for the UK. However, these results can be explained. In the UK, 71% of private school enrolment is socially elitist and most of the schools are expensive and government independent. Thus, these schools have financial constraints and are expected to maximize a profit function. In other words, they are expected to have higher qualities and higher achievements than public schools in order to attract any students (positive sign on the regression coefficient). In contrast, private schools in Finland are government dependent, and are funded and controlled by the state. They are not socially elitist (they are not selective and do not perceive tuition fees) and usually are attended by students who cannot follow the regular curricula in public schools. Hence, they are not expected to perform better than their public counterparts. Similarly, in Japan, the private sector was conceived in order to complement public school supply and is not highly elitist, while in Italy most private schools are catholic non-elitist and non-subsidized. The negative effect that private schooling has on performance scores is a clear indication that the apparent superiority of private schools is channeled through better peer quality or funding and not through structural differences between the two sectors. In other words, when peer effects and funding are controlled for in a regression analysis, the effect of private schooling becomes non-significant or even negative.

9.4.4 The Variance Components

Germany has the highest ratio of between-school variance over total variance, followed by Italy, Japan, the UK and Finland. The high level of between/total ratio in Germany, Italy, and Japan indicates that schools tend to have specific effects that diverge from the average effect (overall intercept). Note that these countries have already streamed their students into differentiated schools. The disparities in Italy may also be the reflection of important territorial differences between the North and the South in addition to stratification in the upper secondary phase. In Germany they reflect differences between general and vocational tracks. In Japan they reflect the hierarchical and stratified nature of the education system in the upper secondary phase which is also differentiated between general and vocational schools.

9.5 Policy Implications

A number of policies were used in different countries to improve equity in the distribution of achievements. These include additional educational resources for particular schools based on their performances and social intakes. Such policies were used in France, (zone d'éducation prioritaire), and in the England, where funds were provided for equalization purposes in favor of poor neighborhoods' schools. Other policies consist of spending more on students presenting specific characteristics, such as belonging to a disadvantaged social class. These policies include vouchers and conditional cash transfers.

On the one hand, policies designed to enhance the situation of individuals should be used in countries were ESCS and other student level variables have a large and significant impact on achievements and on the formation of inequalities (e.g. Germany and the UK). On the other hand, policies designed to enhance the situation of schools should be used in countries were the heterogeneity of schools is the main source of inequalities (e.g. Italy, Germany, and the UK).

Other types of policies that could enhance performance scores for unprivileged social groups are related to the geographical organization of educational supply. Different school choice policies have been used across the OECD countries, ranging from free choice to a strict application of catchment areas. However, a middle solution consists of the use of zoning policies through which district boundaries are fixed in a manner that maximizes achievements and enhances their distribution. This type of policy is supported by my findings. The nonlinearity of peer effects in their means (in the UK, Italy and Japan) suggests that student allocation is not a zero sum game, and that achievements can be enhanced through a better distribution of peers. Hence, the induced reallocation of students can be the tide that lifts all boats.

A final concern would be private schooling and public subsidies to private schools. As my results have shown, after controlling for student and school characteristics, private schooling does not have a positive impact on performance scores except in the UK. Hence, the advantages that private schools may offer are channelled through higher peer quality or higher funding and not through structural differences between public and private schools. Public subsidies to private schools have so far been used to maintain a choice outside the public system. However, such subsidies are subject for debate when private schools become the schooling institutions for the social elite. In this case, the question that can be asked is: why should public subsidies be maintained when the access to private schools is selective and does not favour equality of opportunities? Perhaps the most coherent answer is the one applied in Greece, where private schools exist but are not subsidized. In other words, a school system should offer equal opportunities to all students; yet it should maintain freedom of choice for those who have special tastes in education (religious, etc.), without subsidizing these particular tastes. Moreover, private schools should be subsidized when they provide education to students with particular needs that the public sector cannot satisfy.

9.6 Conclusion

On the one hand, the findings shed light on the mechanisms of stratification and inequalities in attainments. On the other hand, the comparative analysis allowed for a better understanding of the functioning of these mechanisms under different schooling systems.

The results showed clearly that comprehensiveness-driven school homogeneity is a source of equality since it dilutes the impact of schools on performance scores. Moreover, the trade-off between equity and efficiency (high average achievements) does not necessarily exist since Finland combines high levels of achievements with high levels of equity in their distribution. The rest of the countries have higher levels of inequalities than Finland for different reasons. Early selection and the high levels of social disparities in Germany mean that inequalities are transmitted through school and household characteristics. This is also the case in the UK even though inequalities are more moderate. In Italy, household characteristics have limited effects and inequalities are transmitted through school characteristics. This finding reflects school heterogeneity in terms of their funding and peer quality levels. The case of Japan is probably the most ambiguous because Japanese students have been tracked into differentiated high schools 3 months before the PISA tests. Though, what is possible to say is that increased school heterogeneity will definitely lead to higher school-generated inequalities. In general, I can conclude that the delayed selection of students (Finland) is associated with limited and delayed inequalities. Other major findings include the following. Firstly, Private schooling is found to have a negative effect on performance scores in all countries expect in the UK, indicating that the apparent superiority of private schools is the result of better peer quality and funding. Secondly, social peer effects are non-linear in their means in three of the selected countries indicating that the distribution of peers within schools also affects their performances. Finally, it is important to note that this analysis can be further extended through the inclusion of country-level data that accounts for the non-school macro characteristics of each country. Furthermore, inequalities can also be treated as dynamic if the necessary data is available.

References

Arnott, R., & Rowse, J. (1987). Peer group effects and educational attainment. *Journal of Public Economics, 32*, 287–306.

Barzel, Y. (1973). Private schools and public school finance. *The Journal of Political Economy, 81*, 174–186.

De Bartolome, C. (1990). Equilibrium and inefficiency in a community model with peer group effects. *Journal of Political Economy, 98*, 110–133.

Epple, D., Filimon, R., & Romer, T. (1993). Existence of voting and housing equilibrium in a system of communities with property taxes. *Regional Science and Urban Economics, 23*, 585–610.

Epple, D., & Platt, G. (1998). Equilibrium and local redistribution in an urban economy when households differ in both preferences and income. *Journal of Urban Economics, 43*, 23–51.

Epple, D., & Romano, R. (1998). Competition between public and private schools, vouchers, and peer group effects. *The American Economic Review, 88*, 33–62.

Epple, D., & Romano, R. (2006). Admission, tuition, and financial aid policies in the market for higher education. *Econometrica, 74*, 885–928.

Fernandez, R., & Rogerson, R. (1996). Income distribution, communities, and the quality of public education. *The Quarterly Journal of Economics, 111*, 135–164.

Gill, J. (2008). *Bayesian methods: A social and behavioural sciences approach* (2nd ed.). London: Chapman & Hall/CRC.

Green, A. (1997). *Education, globalization, and the nation state*. Basingstoke, UK: Palgrave Macmillan.

Green, A. (1999). East Asian skills formation systems and the challenge of globalization. *Journal of Education and Work, 21*, 253–279.

Green, A., Preston, J., & Janmaat, J. (2006). *Education, equality and social cohesion: A comparative analysis*. London: Palgrave Macmillan.

Hanushek, E., & Welch, F. (2006). *Handbook of the economics of education*. London: Elsevier.

Mostafa, T. (2009). *The anatomy of inequalities in educational achievements: An international investigation of the effects of stratification* (LLAKES Research Paper No. 3). London: Institute of Education.

Nechyba, T. (1997). Existence of equilibrium and stratification in local and hierarchical Tiebout economies with property taxes and voting. *Economic Theory, 10*, 277–304.

Nechyba, T. (2003). Centralization, fiscal federalism, and private school attendance. *International Economic Review, 44*, 179–204.

OECD. (2003a). *PISA data analysis manual for SAS users*. Paris: OECD.

OECD. (2003b). *Education at a glance*. Paris: OECD.

OECD. (2003c). *Learning for tomorrow's world*. Paris: OECD.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). London: Sage Publications.

Robert, C., & Casella, G. (2004). *Monte Carlo statistical methods* (2nd ed.). New York: Springer.

Rose-Ackerman, S. (1979). Market models of local government: Exit, voting, and the land market. *Journal of Urban Economics, 6*, 319–337.

Stiglitz, J. (1974). The demand for education in public and private school systems. *Journal of Public Economics, 3*, 349–385.

Westhoff, F. (1977). Existence of equilibrium in economies with a local public good. *Journal of Economic Theory, 14*, 84–112.

Part IV
Research on Trends in PISA

Introduction: What Are Trends and Why Are They Important for PISA?

Matthias von Davier

Why Do We Need Trends and What Is the Problem?

The particular methods used for establishing a common scale by linking the different PISA cycles are well documented in the technical report from Organisation for Economic Co-operation and Development (Organisation for Economic Co-operation and Development [OECD], 2009). The link is achieved by a simple adjustment of the estimated item difficulties by means of a transformation constant that matches the difficulties of the common items between two adjacent cycles. This transformation of parameters is carried out using the item difficulties from two separate international calibrations of all participating countries combined. This method is commonly used in the statistical analysis of educational data, and it leads to identical results compared to other methods when the statistical model is correct. This method, however, is by no means the strongest possible link between scales in terms of statistical modeling of multiple populations with constrained common parameters (von Davier & von Davier, 2007). If the model used for analysis does not fit the observed data perfectly, different linking methods based on statistically weaker or stronger equality assumptions will indeed lead to some what different results. Like all statistical models, the models used in PISA and other international assessments are, at best, approximations of the truth (Xu & von Davier, 2010). Among the issues that are not modeled explicitly is, for example, the fact that some tasks in the PISA assessment are organized in groups under a common text passage that serves as the item stem (Monseur, Baye, Lafontaine, & Quittre, 2011). Another issue that is not explicitly modeled is that students taking the PISA assessment will, once in a while, omit the response to one or more items. While this may be related to student proficiency, it is by no means clear whether every omitted item indicates with certainty that a certain lack of knowledge or skill is present.

M. von Davier, Research Director
Research & Development Division, Educational Testing Service,
MS 13E, Princeton, NJ 08541, USA
e-mail: mvondavier@ets.org

Model-based approaches to account for these missing responses statistically are available (Glas & Pimentel, 2008; Moustaki & Knott, 2000) and can be adapted to PISA (Rose, von Davier, & Xu, 2010) but are typically not applied in the analysis of PISA data or data from other international assessments.

In order to show how a more rigorous linking methodology can be used with PISA data, Oliveri and von Davier (2011) developed an approach that is based on the strongest possible linking between countries, a method that can be applied to a concurrent analysis of two adjacent cycles treating all common test items as statistically identical. Going from this strongest statistical link between countries, the authors developed a series of statistical models that relax the assumptions made in return for an improved fit of the model to the observed data. This approach could be easily extended to analyses of multiple PISA cycles allowing a concurrent analysis of trend data while improving model data fit at the same time. With regard to the current assessment design and linking methodologies, Mazzeo and von Davier (2008) reviewed the PISA linking design and compared it to the linking methods (Yamamoto & Mazzeo, 1992) and assessment design used in the National Assessment of Educational Progress (NAEP). They concluded that while the more conservative linking design of NAEP appears to result in a more stable link, the linking in PISA is as stable as the current assessment design allows. Adams (2009) concluded that linking in PISA is generally stable. Adams, however, also presented evidence that linking in PISA may be affected for some countries by a fluctuation in the number of link items over cycles because of the design-dependent change between minor and major assessment domains. This change in the number of items for minor and major assessment domains between cycles is unique to the current PISA design and the resulting effects on trends is what led to the research presented in subsequent chapters.

Trends for Major and Minor Domains

PISA, in contrast to other international assessments and to NAEP, employs an assessment design that changes the construct coverage from cycle to cycle. The domains of Reading, Mathematics, and Science rotate in prominence every 3 years so that one is the major domain while the other two are treated as minor domains. and covered in the assessment using fewer item blocks While some of the issues related to this design have been raised by Mazzeo and von Davier (2008) as well as Adams (2009), the three chapters in this section delve into the consequences of alternating between major and minor domains for trend reporting in more detail. More specifically, it became apparent that trend results seem to depend partially on whether the items in the minor and major domain over time were selected in certain ways. If all items in a major domain were used to estimate results for each country, results differed somewhat from those obtained when only the items that are also represented in minor domain cycles were used to measure trend. The rationale for the approach to restrict the trend measure to what is common is based in the adage, “If you want to measure change, do not change the measure!” (Beaton, 1990). Following this advice, one would expect to get a purer measure of trend when basing the trend measure over time on only those items that are common across measurement occasions. Another issue that is discussed in the subsequent contributions is that of the international versus national comparison of trend results. If each country would analyze its data separately, slightly different estimates of item parameters would be obtained. Therefore, trend measures based on national item parameters would be somewhat different from trend measures bases on international parameters. This is based on the fact that common item parameters across countries may not fully reflect that the items administered in different languages most likely do not function exactly the same across countries. In PISA, this is referred to as country-by-item interaction or country differential item functioning (DIF). It is acknowledged that this interaction exists in practically all large scale assessment programs. The recent study by Oliveri and von Davier (2011) proposed a solution for this issue, but common practice so far is to assume international parameters are exactly the same for all participating countries. The fact that country-by-item interactions are studied points to an awareness that international parameters are a compromise at best. The chapters in this part of the volume address the consequences of major and minor domain changes and of country-by-item interactions on trend measurement.

In Chap. 10, "An Investigation of Australian OECD PISA Trend Results," Daniel Urbach discusses a series of different scale linking methods applied to PISA trend data when looking at country specific trend measures for Australia. The author describes the development of country-specific trend scales for Mathematics and Science, and how country-specific results differ somewhat for the Reading domain. Notably, Urbach indicates that trend results reported in the media based on international parameters would have been reported with a somewhat different emphasis if national parameters had been used to report trends.

"Success Despite the Odds? Outcomes for Low-Performing Students in Australia" is Chap. 11 in this volume. In this contribution, Sue Thomson and Kylie Hillman examine low performing students from the cohort followed in the Longitudinal Surveys of Australian Youth (LSAY) study from secondary school to their post-school pathways. This group, part of the population assessed in the PISA 2003 cycle in Australia, was composed students who did not reach proficiency level 3 on the PISA assessment, and the group was followed over the next 4 years. This study is an example of studies that are directly connected to PISA and that rely on the accuracy and stability of the trend data provided by PISA. An interesting feature of this study is that it uses state of the art statistical modeling for multilevel data to connect PISA test results from 2003 and noncognitive measures such as motivation and social competencies to occupational outcomes 4 years later.

Chapter 12, "Linking PISA Competencies over Three Cycles – Results From Germany," is by Claus H. Carstensen. In this contribution, Carstensen discusses some of the limitations of trend analyses based on data from multiple PISA cycles and the effects of the rotating minor vs. major domain definition in the assessment design. A careful analysis of item-by-time-point interaction, a concept analogous to the item-by-country interaction, that looks at changes of item difficulty over cycles, leads to a deeper understanding of the changes over time. Carstensen concludes with the development of a model that utilizes only the short version (the minor domain of) Reading as well as the common items of the minor domains in Mathematics and Science to determine a country specific measure of trend.

All three chapters provide evidence of the importance of providing accurate trend data and how reporting trends based on national versus international estimates of quantities may affect results. These contributions add important evidence to the body of work that will help PISA improve and evolve its assessment design and trend estimation and reporting for future cycles of this important assessment program.

Outlook

In spite of academic discussions about technical issues concerning the improvement of stability and accuracy of trend results, the data made available by PISA and other educational assessments are uniquely useful in assessing where educational systems are headed. The number of data-points available from PISA enables, by the time the ink in this book dries, comparisons of student populations that stretch across more than a decade. The first PISA research conference held in 2009 looked back at three cycles of PISA that have been carried out in 2000, 2003, and 2006. The 2009 data were then in the process of being collected and compiled for analysis and reporting, while the 2012 data will be collected by the time this volume is available.

What can we expect from exploration across five or more cycles of PISA? Why should countries continue to participate in PISA? A trend observed across two or three data points may mean a lot, or may mean not much, given that trend estimates, like all statistics, are associated with uncertainty in the estimate. The more cycles are available, the more certain becomes an estimate of a positive trend (if there is one) or a negative or flat trend. With data from five cycles of PISA available soon, researchers can examine the direction that countries appear to be heading and can make statements that are likely to hold up over the next few rounds (if nothing changes) with more confidence than what can be said about a changes across two or three observations only.

Educational policies will change, and have changed, and will continue to change. In part these changes were implemented as a direct reaction to results from PISA or other available data; in part these changes occurred independently. An ongoing monitoring of educational outcomes over decades will allow an evaluation of these changes in terms of student outcomes. If results from PISA are combined with data from PISA-related longitudinal studies and analyzed with explanatory statistical methods (von Davier, Xu, & Carstensen, 2011) and (cluster) randomized trials are linked to PISA using appropriate statistical models, a host of knowledge can be generated that can be used to inform educational policies. PISA will certainly be an important source of data in itself. In addition, countries have started to design and connect their national assessments to PISA and other studies and to explore ways of how to integrate results.

References

Adams, R. (2009). Trends*: Are they an outrageous fortune or a sea of roubles?* Keynote presented at the 1st PISA Research Conference, University of Kiel, Germany.

Beaton, A. E. (1990). Introduction. In A. E. Beaton & R. Zwick (Eds.), *The effect of changes in the national assessment: Disentangling the NAEP 1985-86 reading anomaly* (Report No. 17-TR-21). Princeton, NJ: National Assessment of Educational Progress.

Glas, C. A. W., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68*, 907–922.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results* (doc.ref. EDU/PISA/GB(2008)28). Retrieved from http://www.oecd.org/dataoecd/44/49/41731967.pdf

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. In M. von Davier & D. Hastedt (Eds.), *IERI monograph series – Issues and methodologies in large scale assessments* (Vol. IV). Hamburg, Germany: IEA/ETS Research Institute (IERI).

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales using latent variable models with covariates. *Journal of the Royal Statistical Society, Series A, 163*(3), 445–459.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315–333.

Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report*. Paris, France: Author.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with IR*T (ETS Research Report No. RR-10-10). Princeton, NJ: ETS.

von Davier, M., & von Davier, A. (2007). A unified approach to IRT scale linkage and scale transformations. *Methodology, 3*, 115–124.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). *Measuring growth in a longitudinal large scale assessment with a general latent variable model.* Psychometrika, online first. DOI: 10.1007/S11336-011-9202-Z.

Xu, X., & von Davier, M. (2010). Linking *errors in trend estimation in large-scale surveys: A case study* (ETS Research Report No.RR-10-11). Princeton, NJ: ETS.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*(2), 155–173.

Chapter 10
An Investigation of Australian OECD PISA Trend Results

Daniel Urbach

Abstract The first three PISA cycles occurred in 2000, 2003 and 2006 with Literacy Scales in Reading, Mathematics and Science. This chapter explores equating-related issues for the Australian data and considers the implications for Australia's reported results. Previous published PISA results have employed a common reporting scale across the first three cycles for Reading only. Common scales for all three Literacy Scales were constructed in this chapter. In addition, the item parameters estimated here were based on Australian data only rather than using the international item parameters, as is done in PISA. This allows for an examination of the impact of country differential item functioning (DIF) on the Australian results. Australian PISA trends were explored in terms of the overall shape of the estimated performance distributions. Where applicable, comparisons were made with the published results based on international item parameters. While such comparisons showed several similarities, some differences were also found.

Published Australian Reading distributions reported a decline over the first three cycles in the performance of Australian students located at the top end of the distribution. Using Australian data only, a decline between the first two PISA cycles was found, but remarkably in the bottom 15% of the distribution only. Between cycles 2003 and 2006 an almost constant decline across the whole proficiency distribution was found and not a decline that was limited to the top end of the distribution, as published by the media.

Reported PISA results have a high impact on educational policy. The outcomes of trend analyses may alter with different methods. This investigation examines the impact when Australian country specific and when International item parameters are used to estimate the distributions of Australian PISA performance. This is further explored by equating the first three PISA cycles for each literacy scale.

D. Urbach (✉)
Psychometrics and Methodology Australian Council for Educational Research,
19 Prospect Hill Road, Camberwell, VIC 3124, Australia
e-mail: daniel.urbach@acer.edu.au

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_10,

The results reported in this chapter highlight some of the potentially important differences that can occur when using different analyses methods.

Keywords Differential item functioning • Pyschometics • Items response theory • Test equation

10.1 Introduction

In past PISA cycles item parameters are set at an international level and consequently the item parameters are the same for all countries within a PISA cycle. One of the main reasons a country may participate in PISA is to monitor its proficiency levels over time. The presence of differential item functioning (DIF) between countries or item by country interaction is not taken into consideration when using a single international item parameter set. Country DIF may be due to various factors such as the translation of an item between languages as well as cultural differences, which may make the same item on paper function differently between countries. Existence of such DIF may influence the soundness of valid international comparisons, and therefore its investigation may be critical.

Evidence of such country DIF is regularly observed in international studies (Adams & Carstensen, 2002; Adams, Wu, & Macaskill, 1998; Mullis & Martin, 1998). Gebhardt and Adams (2007) investigate this issue for Reading and Science across PISA 2000 and PISA 2003. For Australia they find no statistically significant trend differences (after estimating country specific item parameters) between these two PISA cycles for these two PISA domains, except for a small decline in Science when using conditional trends.

Another trend related topic is equating between PISA cycles. After the first three PISA cycles, trends across all three cycles within all three domains have not been publicly reported on. PISA Mathematics is only publicly equated between cycles 2003 and 2006 and PISA Science 2006 is not equated to any of the two previous cycles. However, link items do exist between the three cycles within all PISA literacy scales to allow results to be reported on across the first three cycles within each literacy scale.

The investigations of trends in PISA have important implications with many directly and indirectly associated stakeholders involved with PISA such as Educational, Social and Economic researchers, policy makers, educational systems, students, student parents and teachers. Changes in the observed PISA proficiency distributions over time are widely referred to both in national and international media and hence are highly emphasized publicly.

Various Australian newspapers (Buckingham, 2008; Gale, 2008; Milburn, 2008) have pointed out declining trends in Australian PISA reading performance. They are attributing these declines to a drop in achievement of the highest performing students (based on published results). Hence, they suggest educational policy should focus on top students more and concentrate less on minimum standards and the bottom end of student abilities. Thomson and De Bortoli (2008) in the Australian PISA

National Report, conclude that while Australian student performance levels are well above the OECD average, they are in general, not improving.

Such media examples and the results from the Australian PISA Report further highlight the importance of investigating the impact of estimating Australian PISA trend results using different methods. The aim of this investigation is to examine the impact DIF has on the distributions of Australian PISA performance over time. This is achieved by exploring the differences when Australian country specific item parameters and when international item parameters are used to estimate the distributions of Australian performance. The results produced from equating the first three PISA cycles (2000, 2003 and 2006) for each literacy scale are also explored. This involves equating the minor PISA domains with the major PISA domain over the three cycles, for each literacy scale.

10.2 Methodology

Each cycle of PISA assesses three areas, namely Reading, Mathematics and Science. In each cycle one of these areas is assessed as the major domain and the other two as minor domains. The major domain is given greater emphasis and is assessed with more items than the other two. The first PISA survey took place in 2000 with Reading as the major domain. The second and third PISA cycles took place in 2003 and 2006, with Mathematics and Science as the major domains for these respective cycles. More information on the PISA cycles can be found in PISA technical reports (Adams & Wu, 2002; OECD, 2005, 2009).

10.2.1 Item Calibration

The model used in PISA to calibrate items, is a generalized form of the Rasch Model, which uses a conditional item response model in conjunction with a multivariate population model (Adams & Wu, 2002; OECD, 2005, 2009). The estimations in PISA were made in ConQuest (Wu, Adams, Wilson, & Haldane, 2007). Student abilities were estimated with item parameters anchored at their international calibration values and using the plausible values (PVs) technique developed by Mislevy (1991).

In each PISA cycle, items were calibrated at the international level, giving the same item parameters for every country. For the international item calibration, a sub-sample of students referred to as the international calibration sample, was taken, which comprises 500 students drawn at random from each of the participating OECD countries that met PISA response rate standards. In PISA 2000 the international calibration sample consisted of 13,500 students in 27 participating OECD countries. In PISA 2003 and in PISA 2006 the international calibration sample consisted of 15,000 students in 30 participating OECD countries.

Table 10.1 Summary of major PISA cognitive reporting scales

Name	Established	2000	2003	2006	Comment
PISA literacy scale					
PISA Reading	2000	✓	✓	✓	Trends can be reported between any of the three cycles, by country or by subgroups within countries
PISA Mathematics	2003		✓	✓	Trends can be reported between 2003 and 2006, by country or by subgroups within countries
PISA Science	2006			✓	Provides the basis for future trend analysis by country or by subgroups within country

In this chapter, the calibration method has the item parameters set at the country level using the Australian data only, giving country specific item parameters for Australia. Setting Australian specific item parameters allows for an examination of differential item functioning (DIF) on the Australian results, as any existing DIF due to all other countries is removed. The rest of the analysis used the same methods as was done for the published results.

10.2.2 PISA Equating Designs

For PISA 2000, 2003 and 2006, scales in Reading, Mathematics and Science Literacy have been reported on. Over the first three cycles of PISA, a total of 19 distinct scales have been produced and published. These include the PISA Literacy Scales, PISA Literacy Subscales (sub-components of the literacy scales domains) and Special Purpose Scales (additional and provisional scales). The focus of this investigation is to examine the major scales only, i.e. the PISA Literacy Scales. Table 10.1, taken from the Scaling Outcomes Chapter in the PISA 2006 Technical Report (OECD, 2009), shows a brief summary of the established literacy scales, and to which cycles they are publicly equated to.

While the PISA Mathematics Literacy Scale in 2000 is not equated in the international analysis to the following two cycles (as it only became a major domain in 2003), there were common items between all cycles and this makes it possible to explore and equate the Australian data over the first three cycles. Likewise, the international analysis of the PISA Science Literacy scales in 2000 and 2003 are not equated to the 2006 scale (as it only became a major domain in 2006). However, common items exist between the first three cycles, again allowing for the exploration and equating of the three cycles for the Science domain.

Scales were constructed using Item Response Theory (IRT) methods (Lord, 1980) and implementing the mixed coefficients multinomial logit model (MCMLM; Adams, Wilson, & Wang, 1997) in ACER ConQuest (Wu et al., 2007). In addition, common item equating and the concurrent analysis methods (Baker, 1984; Lord, 1975), allowed for scores on different tests with varying degrees of

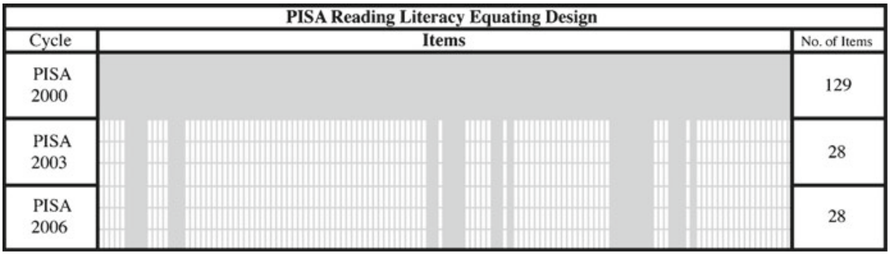


Fig. 10.1 PISA Reading literacy equating design

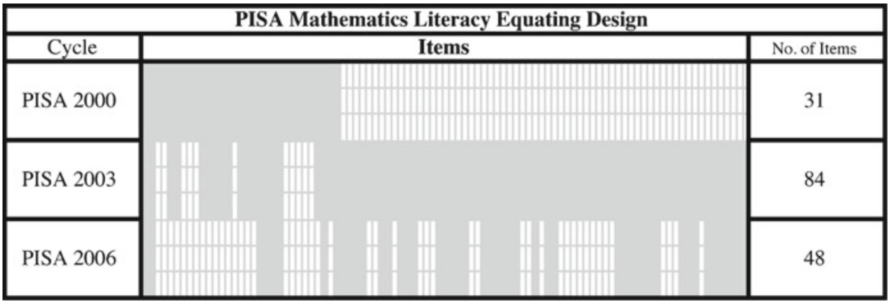


Fig. 10.2 PISA Mathematics literacy equating design

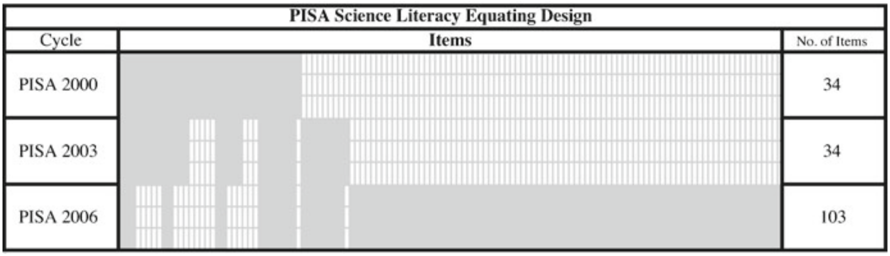


Fig. 10.3 PISA Science literacy equating design

difficulties to be made comparable by construction of a common measurement scale. Such an equating methodology was employed through horizontal item equating, linking between booklets within a cycle and also linking across each cycle. This was completed for the first three PISA cycles. Distributions based on the concurrent analyses were reported on in log units (or logits). It is worth pointing out that published distributions are reported in Pisa Scale Scores which are just a transformation from the international logits. However, the international logits and the logits estimated here, do not have the same unit length and can hence only be indicatively compared.

Figures 10.1, 10.2 and 10.3 show the equating designs for the Reading, Mathematics and Science Literacy scales over the first three cycles. The cycle with

Table 10.2 Number of link items between successive PISA assessments

	Reading	Mathematics	Science
As major domain	129	84	103
Link 2000–2003	28	20	25
Link 2003–2006	28	48	22

the most items in each of the three PISA Literacy scales, is the cycle in which that Literacy scale was the major domain (i.e. Reading in 2000, Mathematics in 2003 and Science in 2006).

As shown in Fig. 10.1, to link the PISA 2000 Reading Literacy Scale with the PISA 2003 Reading Literacy scale, 28 common items were placed in both cycles. Between 2003 and 2006 the Reading test was the same.

For PISA Mathematics Literacy, 20 of the 31 items in 2000 were common items to 2003. This is shown schematically in Fig. 10.2. All 48 items in 2006 are common to 2003, while 8 of these common items also link between 2000 and 2006.

For PISA Science Literacy, 25 of the 34 items in 2000 are common to 2003. In 2003, 22 of the 34 items are common items with 2006. Between 2000 and 2006 there are 14 common items. This is again shown schematically in Fig. 10.3.

Table 10.2, also taken from the Scaling Outcomes Chapter in the PISA 2006 Technical Report (OECD, 2009), summarises the equating designs.

Using the Australian data only, concurrently equated data sets were set up for the first three PISA cycles for each of the three Literacy domains. With these data files, Australian item parameters were calibrated and single scales were constructed for each Literacy domain. This was carried out using a unidimensional concurrent analysis and conditioning on each PISA cycle (using PISA 2000 as the reference category). The proficiency distributions obtained after performing these calibrations, were then used to estimate Australian data based trend results.

Differences in the trends across the proficiency distributions are investigated in the next section. This is done for the distributions of Australian performance, based on the Australian data item calibrations and, for the published Australian distributions which are based on the international item calibrations.

10.3 Australian Distribution Trends: Percentiles

10.3.1 Australian Reading Percentile Trends

Differences between PISA Reading Literacy trends were investigated, by examining the distributions of Australian Reading proficiency. This was achieved after conditioning on the first three cycles (using PISA 2000 as the reference). The percentiles of the Reading Literacy distribution were calculated and represented in the form of de-trended Q-Q (Quantile-Quantile) plots. These plots simply show the differences between cycles at each percentile point. Such plots allow for pair wise comparisons

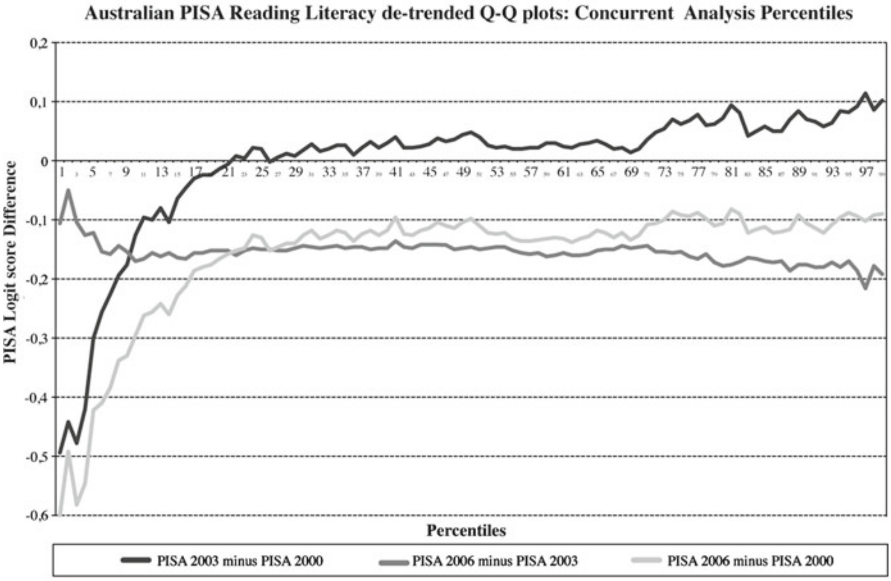


Fig. 10.4 De-trended Q-Q plots of concurrent Australian PISA Reading cycles (logits)

(between cycles in this case) of the proficiency distributions, making it possible to view specific points and ranges of the distribution which are causing differences in the trends. Figure 10.4 compares the percentile distributions of the Australian PISA Reading proficiencies in PISA 2003 with PISA 2000, PISA 2006 with PISA 2003 as well as the differences in the distribution over the three cycles between PISA 2006 and PISA 2000, based on the concurrent analyses.

When there are no differences between percentile points across the distribution, the distributions have both the same location and shape (i.e. they are the same). When the location of the plot is positive, there is an increase in performance and when the plot is negative, there is a decrease in performance over the two particular cycles.

The 2003 and 2000 distributions are shown to have very similar shapes and locations in the middle of the percentile range. Very notable however, is the sharp decline in the bottom 15% of students from cycle 2000 to 2003, by an average of around a quarter of a logit. There is also a small increase at the top of the proficiency distribution, from cycle 2000 to cycle 2003, from around the 70th percentile upwards.

PISA Reading Literacy between cycles 2003 and 2006 displays an almost constant decline in proficiency across the whole distribution. The decline for the bottom 75% of students hovers around 0.15 of a logit. This decline approaches 0.20 of a logit for the top 25% of students. The overall mean difference between these two distributions is also statistically significant at the 5% level with an overall mean decline of 0.16 logits.

The de-trended Q-Q plot between 2006 and 2000 is simply the sum of the previous two plots and highlights the changes over the three cycles. A mean decline of 0.33 logits is found between the 2000 and 2006 cycles for the bottom 20% of students. The rest of the distribution displays a steady decline by an average of 0.12 logits.

To put the magnitude of the differences found into perspective it is worth pointing out that the width of each published PISA Reading band is 0.80 logits (however such a comparison is only indicative and assumes the unit length to be the same in both the concurrent and published scales). The same band width comparison applies to Mathematics and Science below.

10.3.2 Australian Published Reading Percentile Trends

Internationally, there are 36 countries in which Reading Literacy performance can be compared between PISA 2000 and PISA 2006. The Australian National PISA report states that in PISA 2000 only Finland outperformed Australia. In 2003 Australia was outperformed by Finland and Korea, while in PISA 2006 five countries outperformed Australia by statistically significant reading margins (Finland, Korea, Hong Kong – China, Canada and New Zealand).

When the Reading Literacy scale was constructed in PISA 2000, the resultant scores were set to have a mean of 500 PISA scale score (PSS) points and a standard deviation of 100 PSS points, across the participating OECD countries. This OECD mean declined to 494 PSS points in PISA 2003 and further down to 492 PSS points in PISA 2006.

The distributions from the published Reading Analyses are compared across cycles in Fig. 10.5. The notable difference here, compared to the concurrent distributions, is that the published distributions do not show the sharp decline in the bottom 15% of students from cycle 2000 to 2003, but instead, display a reasonably similar shape and distribution location between the two cycles up until around percentile 70 (with a slight rise in the lower end and a slight fall in higher end of the bottom 70 percentile points). At the top end of the distribution another difference is found. PISA 2003 estimated proficiencies become increasingly lower than the PISA 2000 published Reading Literacy results. Between percentiles in the 70s, 80s and 90s, the average decline is around 8, 11 and 13 PSS points respectively. It is noted here, however, that any comparisons of the concurrent and published distributions are limited, as the units of the plotted distributions from the concurrent analyses are in logits and the units of the published distributions shown here are in PISA scale scores. While the units are different, the scales and students are not and hence still provide very useful insight.

When comparing the 2006 and 2003 Reading Literacy distributions a decline across the distribution is found. The higher ability groups have declined the most. The published PISA Reading Literacy scores of the bottom 20%, the 21st to 40th, the middle 20% (percentile 41–60), the 61st to 80th percentile range, and the top

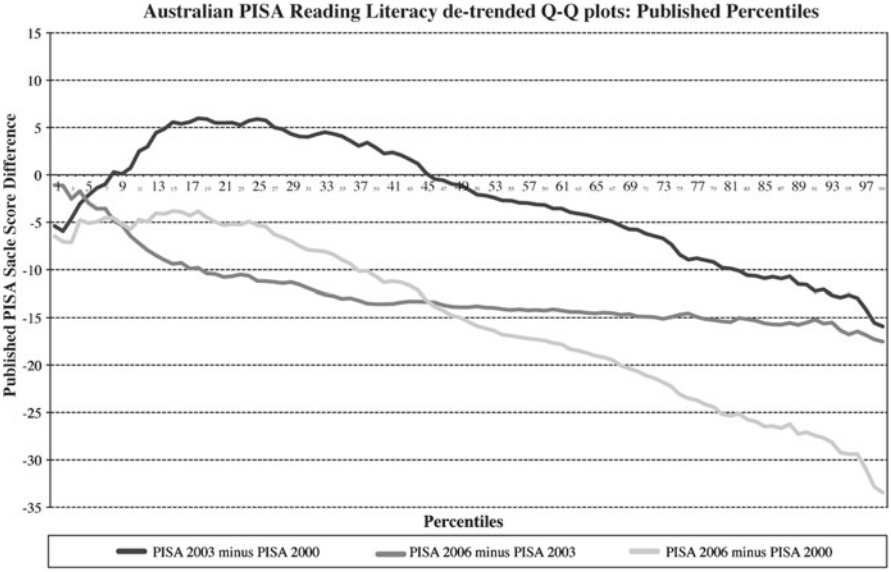


Fig. 10.5 De-trended Q-Q plots of published Australian PISA Reading cycles (PSS)

20% of students, decline by an average of 6, 12, 14, 15 and 16 scale score points respectively, showing a constant decline from the middle of the distribution onwards. This slightly contrasts the concurrent analysis results in Fig. 10.4, which demonstrates an almost constant decline in proficiency across the whole distribution, from cycle 2003 to cycle 2006. Overall, the mean difference between these two distributions is also statistically significant at the 5% level with an overall mean decline of 12.4 PSS points.

The decline in published PISA Reading Literacy is even more evident when comparing cycles 2006 and 2000, showing the decline over the three cycles. Particularly at percentile 40 and higher, where proficiency levels in PISA 2006 compared to PISA 2000 continue to further decrease as percentiles increase.

Over the three PISA cycles, the Reading Literacy scores of the bottom 20% and the 21st to 40th percentile ranges declined by an average of 5 and 8 scale score points respectively. The middle 20% (percentile 41–60) and the 61st to 80th percentile range declined by an average of 15 and 21 scale score points respectively. The top 20% of students, decline by an average of 28 scale score points. This decline is especially large for the top 40% of students. At the 99th percentile the decline is 33 scale score points over the three cycles.

In contrast, the decline reported in the concurrent results, does not increase as the percentiles increase but tends to be constant across the proficiency distribution, with the sharp decline for the bottom 15% of students.

To again give these differences some reference, it is worth pointing out that the size of a standard deviation is around 100 scale scores.

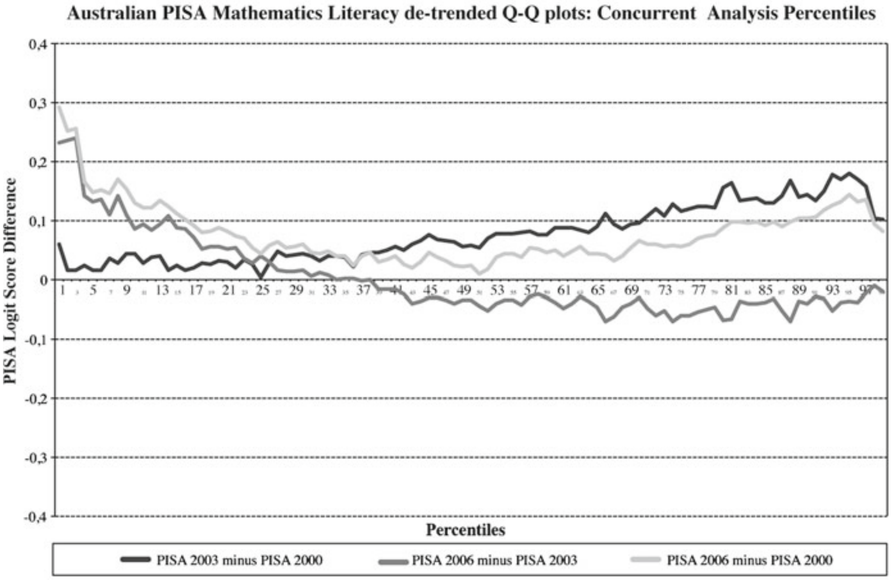


Fig. 10.6 De-trended Q-Q plots of concurrent Australian PISA Mathematics cycles (logits)

10.3.3 Australian Mathematics Percentile Trends

Figure 10.6 compares the cycles of the distributions from the Concurrent Mathematics Analyses. The distributions between cycle 2000 and 2003 are very close to each other with some increases in the top students. From around the 35th percentile onwards, there is a slight increase between cycle 2000 and 2003 which continues to gradually rise all the way to the top end of the distribution. The average increase from cycle 2000 to cycle 2003 for the 61st to 80th percentiles is 0.11 logits and the top 20% of students Mathematics ability increase on average by 0.15 logits.

The comparisons of the 2003 and 2006 Concurrent Mathematics Literacy distributions indicates that for low percentile points (the bottom 20 percentiles points) 2006 performance is slightly higher than in 2003 (by an average of 0.12 logits). The rest of the distribution displays similar performance in both cycles with a very small and steady decline in the top 65 percentile points.

Over the three cycles, the slight rise between 2006 and 2003 is found in the lower part of the distribution, and the slight rise between 2003 and 2000 is found in the upper part of the distribution. The middle of the distribution shows little to no change. The average increase in the bottom 20% is 0.15 logits and the top 20% rises by an average of 0.11 logits.

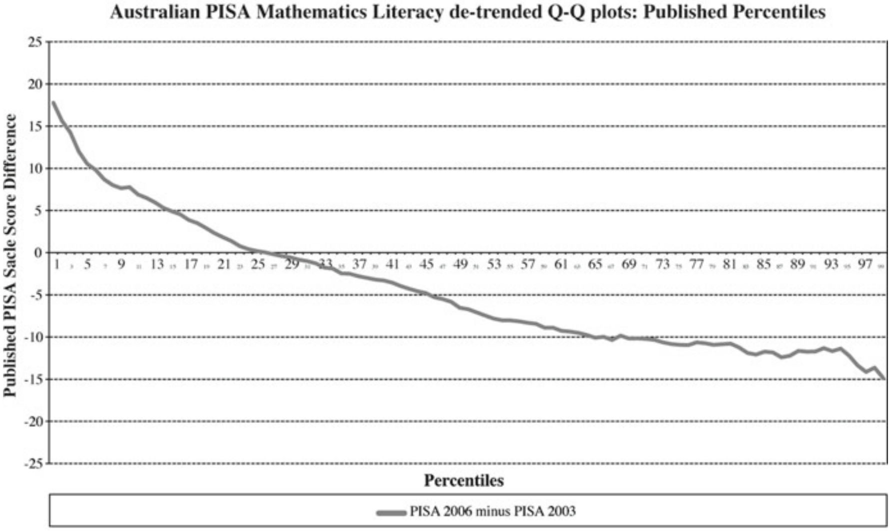


Fig. 10.7 De-trended Q-Q plots of published Australian PISA Mathematics cycles (PSS)

10.3.4 *Australian Published Mathematics Percentile Trends*

Internationally, there are 39 countries in which Mathematical Literacy performance can be compared between PISA 2003 and PISA 2006. The Australian National PISA report states that in PISA 2006, eight countries significantly outperformed Australia. These were Chinese Taipei, Finland, Hong Kong-China, Korea, the Netherlands, Switzerland, Canada and Macao-China.

In 2003, when Mathematical Literacy was a major PISA domain, the scores from the constructed scale were also set to have a mean of 500 PSS points and a standard deviation of 100 PSS points, across the participating OECD countries. The OECD mean declined only slightly to 498 in PISA 2006.

Figure 10.7 shows the distribution comparison of the only two comparable cycles published for PISA Mathematics (2003 and 2006).

The comparison shows that for low percentile points (around the bottom 20%) 2006 performance is slightly higher than in 2003. This is consistent with the results of the concurrent percentile distributions for these two cycles. The next part of the distribution (between the 21st and 40th percentile points) displays similar performance in both cycles. Similar to Reading there is actually a decline in the top 60 percentile points in 2006. Although the published results are in PISA Scale Scores, a similar pattern (but weaker in terms of the decline) is found for these two cycles compared to the concurrent analysis results shown above.

The published PISA Mathematics Literacy scores of the bottom 20% of students increased by an average of 8 scale score points over the two cycles, student scores

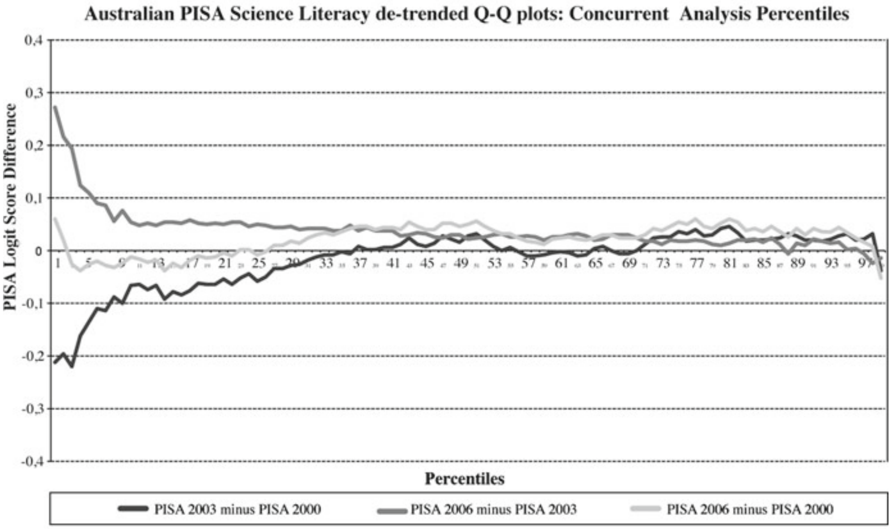


Fig. 10.8 De-trended Q-Q plots of concurrent Australian PISA Science cycles (logits)

between the 21st to 40th percentile declined by an average of only 1 scale score (which was not statistically significant), the middle 20% (percentile 41–60) declined by an average of 7, the 61st to 80th percentile and the top 20% of student declined by an average of 10 and 12 scale scores respectively. This decline is not as large as it was found for Reading, although it was statistically significant for the top 40% of students. At the 99th percentile point the decline is 15 scale score points over the two cycles.

10.3.5 Australian Science Percentile Trends

After investigating Australian PISA Reading and Mathematics distributions, Australian PISA Science distributions were also explored.

While plots for published PISA Science results could not be produced as results from cycle 2006 were not publicly equated back to the previous two cycles, this was done for the Concurrent Science Analysis.

As shown in Fig. 10.8, the distributions between 2000 and 2003 are very close to each other apart from the bottom 20% of students which show an average decline of around 0.10 logits.

Similarly there is also little to no change between the 2003 and 2006 cycles, apart from a slight increase in the bottom 20% of students (by an average of 0.08 logits).

Changes in the bottom of the distribution average out over the three cycles and hence there is little to no change across the three cycles for the concurrent analysis of the Australian Science distribution.

After the first three PISA cycles, only some Science trends for PISA were published (PISA 2006 Report, Annex A7 (OECD, 2007)), which were means in 2003 and 2006 estimated using link items only. Using Science link items only, the Australian 2003 and 2006 cycles were reported to have means of 529.9 and 529.2 PSS points and standard errors of 3.1 and 2.2 PSS points respectively. The distribution comparisons made here tends to agree with this steady trend from 2003 to 2006.

10.4 Concluding Observations

PISA equating and distributions of performance were explored with a focus on Australian data over the first three PISA cycles (2000, 2003 and 2006). Scales were constructed for each PISA literacy scale (Reading, Mathematics and Science) with Australian country specific item estimates (previous PISA calibrations are based on international item estimates).

Australian PISA results were investigated by making pair wise comparisons (between cycles) of the proficiency distributions at each percentile point. This was done using the results based on the Australian specific item parameters as well as the published results.

As reported in the media, published Australian Reading distributions showed a decline in achievement at the top end of the distribution. Between cycles 2000 and 2003 there was a decline from around the 70th percentile onwards and between cycles 2003 and 2006, the decline was even more severe; the higher the ability group the higher the decline from around the 20th percentile onwards.

In contrast to published results, trends of the Australian Reading distributions estimated using Australian data only, showed a decline between the first two cycles, but remarkably in the bottom 15% of the distribution and between cycles 2003 and 2006 an almost constant decline across the whole proficiency distribution. These results contrast the published PISA Australian Reading results and highlight some the differences that can occur when different analysis methods are used. The reason for this could be due to country DIF. The cause could also be partly due to a 2003 PISA rescale, which changed the standard deviations. In 2003, a rescaling was done on the 2000 data with 2003 values anchored for Reading and Science (OECD, 2005). Exploring the cause of the distribution difference certainly warrants further investigation. While Australian educational policies may focus on the improvement of the top end of the Reading distribution due to the published results, the concurrent results suggest the focus of such policies should be targeted towards the bottom end of the Australian Reading distribution.

Published Australian Mathematics results between cycles 2003 and 2006 showed an increase for the bottom 20% of students and a decline for around the top 50 percentile points. In comparison, the results found using Australian Mathematics data only also showed a rise in the bottom 20 percentile points but only a very slight decline in the top 50 percentile points between cycles 2003 and 2006. Between cycles 2000 and 2003, little to no change was found in the bottom 40% of the distribution and in cycle 2003 an increase was found in the remainder of the distribution.

The concurrent analysis also gave some insight into Australian PISA Science trends, which have not previously been published. Only some Science trends for PISA have been published (PISA 2006 Report, Annex A7 (OECD, 2007)), which were means in 2003 and 2006 estimated using link items only. The results shown in this chapter indicate little change across the distribution in Science over the first three PISA cycles.

10.5 Discussion

Distributions of performance across cycles varied between using Australian item estimates and using International item estimates. This indicates the existence of country DIF. The presence of country DIF may be accounted for by allowing items which display DIF to have country specific item parameters. However, this may not be acceptable in practice. Apart from becoming an extremely difficult and tedious task, there would also be political implications. There were over 400,000 students in nearly 60 participating countries in PISA 2006. Some of these participating countries may not accept the use of country DIF adjustments. Especially if a country's results were to be scaled down due to the existence of country DIF.

In any case, discrepancies found in the distributions of the Australian performance, using different methods may seem alarming. It is worth pointing out however, that the aim here was not to imply that the official results need revising but rather to investigate differences that can occur when different methods are used. Therefore, the differences found here should not be treated as alarming but as a reason to keep investigating the impact of country DIF and the implications of this impact.

The equating designs between minor PISA domains and major PISA domains all contain common items for each PISA literacy scale. As was shown here, it would be feasible to equate each literacy scale over the first three cycles and beyond. This would allow the influence on reported PISA results to be explored further for all countries.

References

Adams, R. J., & Carstensen, C. (2002). Scaling outcomes. In R. J. Adams & M. Wu (Eds.), *Programme for international student assessment: PISA 2000 technical report* (pp. 149–162). Paris: OECD.

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

Adams, R. J., & Wu, M. L. (2002). *Programme for international student assessment: PISA 2000 technical report*. Paris: OECD.

Adams, R. J., Wu, M. L., & Macaskill, G. (1998). Scaling methodology and procedures for the mathematics and science scales. In M. O. Martin & D. Kelly (Eds.), *TIMSS technical report* (Implementation and analysis (primary and middle school years), Vol. II, pp. 147–174). Boston: Boston College.

Baker, F. (1984). Ability metric transformations involved in vertical equating under item response theory. *Applied Psychological Measurement, 8*(3), 261–271.

Buckingham, J. (2008, November 6). Brightest and best miss out. *The Australian*. Retrieved from http:// www.theaustralian.com.au.

Gale, T. (2008, August 13). Fair go must no be just a phrase. *The Australian*. Retrieved from http:// www.theaustralian.com.au.

Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement, 8*(3), 305–322.

Lord, F.M. (1975). *A survey of equating methods based upon item characteristic curve theory* (RB-75–13). Princeton NJ: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Milburn, C. (2008, September 15). Literacy fall starts with out best. *The Age, Education*, p. 3.

Mislevy, R. J. (1991). *Randomization-based inference about latent variable from complex samples, Psychometrika 56* (pp. 177–196). Greensboro, NC: Psychometric Society.

Mullis, I. V. S., & Martin, M. (1998). Item analysis and review. In M. O. Martin & D. Kelly (Eds.), *TIMSS technical report* (Implementation and analysis (Primary and middle school years), Vol. II, pp. 111–146). Boston: Boston College.

OECD. (2005). *PISA 2003 technical report*. Paris: OECD.

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

OECD. (2009). *PISA 2006 technical report*. Paris: OECD.

Thomson, S., & De Bortoli, L. (2008). *Exploring scientific literacy: How Australia measures up, PISA national reports*. Camberwell, Australia: Australian Council for Educational Research.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ACER ConQuest version 2.0 [Computer program]*. Camberwell, Australia: ACER Press, Australian Council for Education Research.

Chapter 11
Success Despite the Odds? Outcomes for Low-Performing Students in Australia

Sue Thomson and Kylie Hillman

Abstract This chapter reports on a project that utilised data from the Longitudinal Surveys of Australian Youth (LSAY) study which followed the PISA 2003 cohort from secondary school along their post-school pathways. This project investigated what differentiates low-performing students who have positive and successful outcomes in the years after leaving school from those who have less successful outcomes.

Members of the LSAY Y03 cohort who did not reach Proficiency Level 3 on the Mathematics literacy component of PISA made up the target sample for this project. Their pathways over the subsequent 4 years, through school and into further education, training and the labour force, were tracked, and those who had successful outcomes by age 19 were identified. Similar analyses were also conducted with a sample of high performers, those who achieved at Proficiency Level 5 or 6 on the mathematics assessment, to provide a comparison with the results for the low performers.

Multilevel logistic regression analyses were carried out to identify those factors that distinguished between low performers who 'succeeded', in that they were fully engaged in study, training, employment or a combination of these activities and were happy with various aspects of their lives, and those with less positive outcomes. The sample included 1,596 students from 294 schools. Those who were from high or medium socioeconomic backgrounds, who were scored high on the instrumental motivation scale, enjoyed school, got along well with their teachers, planned to undertake an apprenticeship or came from schools in non-metropolitan areas were more likely to be successful than other sample members, while those who did not have any plans for what they might do after leaving school, particularly girls without firm career or study plans, were less likely to have a successful outcome.

S. Thomson (✉) • K. Hillman
Educational Monitoring and Research, Australian Council for Educational Research (ACER),
19 Prospect Hill Road, Camberwell, VIC 3124, Australia
e-mail: sue.thomson@acer.edu.au; kylie.hillman@acer.edu.au

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_11,

The comparisons made with the outcomes for low and high-performing students highlighted the importance of a positive and supportive school climate, in which all students can find a teacher with whom they can get along. Ensuring that the school experience is a positive one not only impacts on young people's lives at the time they are at school but appears to continue to influence them once they have left.

Keywords PISA • Outcomes • Low performers

11.1 Introduction

The overall aim of the Programme for International Student Assessment (PISA) is to measure how well 15-year-olds (which in most OECD countries corresponds to when young people are approaching the end of compulsory schooling) are prepared for meeting the challenges they will face in their lives beyond school. The OECD has defined level 2 on the PISA proficiency scales as representing a baseline level of literacy at which students begin to demonstrate the competencies that will enable them to actively participate in life situations. Students performing below this baseline, it is argued, are at serious risk of not being able to adequately participate in the twenty-first century workforce and contribute as productive citizens (see, for example, OECD, 2004). In Australia, however, the Ministerial Council on Education, Employment, Training and Youth Affairs (MCEETYA) determined that "the national standards … should be set at a 'proficient' standard, rather than a 'minimum' standard" (MCEETYA, 2006, p. 4), and set the key performance measure as the percentage of students performing at or above Proficiency Level 3 on each of the OECD PISA literacy scales. In terms of assessing whether or not Level 2 or 3 on the PISA proficiency scales does inhibit future outcomes, however, PISA is limited by its cross-sectional design.

So, does it really matter that a young person didn't achieve well, once he or she is in the labour force? Longitudinal research in Australia suggests that it does. Research conducted using data from the Longitudinal Surveys of Australian Youth (LSAY) in Australia has reported a strong relationship between achievement by Year 9 and school completion and participation in post-secondary education and training (e.g. Fullarton, Walker, Ainley, & Hillman, 2003). Low achievers are more likely as a group to leave school early and thus more likely to enter the labour market without a Year 12 certificate or further qualifications (Curtis & McMillan, 2008; Fullarton et al., 2003), while other studies have found that low achievers were at greater risk of unemployment than their higher achieving peers regardless of school completion status (McMillan & Marks, 2003). High achievers, on the other hand, are more likely to make use of opportunities for further education and training, particularly at universities. Higher achievement was found to be associated with higher occupational status for school completers and early leavers alike. The relationships between achievement levels and labour market outcomes remained even controlling for differences in socioeconomic background and other related factors, indicating that lower levels of achievement in areas such as reading and mathematics can exert an enduring influence on the lives of young people.

Nevertheless, this relationship is not always so simple; not all low achievers leave school without completing Year 12, indeed many continue with their education and training at TAFE or university and go on to stable employment. Using the Youth in Transition Study (YITS) data, Thiessen (2007) found that "substantial minorities of young people graduate from high school and participate in PSE (post-secondary education) despite weak earlier academic performance" (p. 1) and refer to this phenomenon as 'educational resilience'. Educational resilience can result from individual-level factors, such as having supportive family and friends who value education, positive attitudes towards education, plans for future study and employment options, or involvement in enriching activities, or system-level factors, such as school climate or high quality teaching, that influence identifiable groups of students (Fullarton, 2002; Khoo & Ainley, 2005; Thiessen, 2007).

This chapter reports on a project that investigated what individual and system level factors distinguish between low performers who go on to have successful outcomes after leaving school from those who have less positive outcomes. The definition of "success" used in this project included satisfaction with life, as well as whether young people were fully occupied with education, employment or a combination of these activities. Those who were fully engaged and happy with their lives were designated as having a 'successful outcome'. These findings were then compared with similar analyses conducted with a sample of high-performing students.

11.2 Longitudinal Surveys of Australian Youth

In 2003, Australia drew a sample of students for PISA that was very much larger than the basic international requirement. One of the reasons for drawing the larger sample was that the PISA 2003 students would become a commencing cohort for the Longitudinal Surveys of Australian Youth (LSAY-known as the Y03 cohort[1]). A large sample is needed to allow for attrition: over time a proportion of the original sample is not able to be traced.

LSAY is a series of surveys (beginning in 1995 but linked to an older longitudinal study) that focuses on the progress of young Australians as they move from their mid-teens to their mid-twenties, from their initial education to independent working life. These surveys involve large nationally representative samples of young people from whom data are collected each year about education and training, work and social development. Data from LSAY surveys provide descriptions of what young Australians are doing as they negotiate the transition from school, document changes as the group gets older, and enable comparisons with other groups when they were

[1] Similarly, the PISA samples for 2006 and 2009 have also formed the commencing cohort for successive waves of LSAY.

the same age. It is envisaged that the link between LSAY and PISA will provide a basis for investigating the enduring effects of the skills and knowledge measured in PISA.

The LSAY data provide a unique opportunity to investigate the pathways of young people who scored poorly on the PISA mathematics tests in 2003 in the later years of secondary school, and to relate their outcomes to other variables, particularly sociodemographic background variables, gender and interests as measured in PISA. Longitudinal data enable the detailed mapping of individual pathways as well as facilitating causal analyses. The PISA data provides a wealth of information not only about student level factors influencing achievement, but also about school-level influences such as school-level perception of school climate and resourcing.

11.3 'Low' and 'High' Performing Students – The Sample

The major focus of the PISA 2003 assessment was mathematical literacy, and the sample of students chosen for the first analysis were those students who did not achieve at least Proficiency Level 3 in mathematics in the PISA 2003 assessment. The decision to examine those students who did not achieve Proficiency Level 3 was taken to bring the project in line with national definitions of groups of concern in education (e.g. those not meeting the national standards for performance).

In addition, a sample of high performers, those who achieved at Proficiency Levels 5 or 6 on the mathematics assessment, were also selected to provided a comparison with the results for the low performers.

11.3.1 Attrition Over Time

Differential attrition of particular groups of respondents over the course of longitudinal studies, through non-response to contact, can lead to bias in analyses of the survey data, as the sample is no longer representative of the original population. A comparison of the original (2003) and retained (2007) samples of low performers in the PISA mathematics assessment revealed a number of differences between those who remained in the study and those who dropped out, particularly in relation to mathematical literacy proficiency levels and aspirations for the future. Young people who performed at Proficiency Level 1 were less likely to remain in the study, suggesting that the results for subsequent analyses may not hold for these lowest performers. For the aspirations variables, however, the differential attrition may act in a way as to dilute the effect of findings: in other words, if the 'lost' students were retained in the sample, it is likely that there would be a stronger effect for the significant factors identified in the analyses.

The multilevel analyses reported here were restricted to the subset of low performers for whom full data were available for the years 2003–2007, and left unweighted,[2] with the acknowledgement that the results of the analyses may not be representative of the situation for the lowest mathematics performers

11.4 Defining a 'Successful' Outcome

A key feature of this study was the multifaceted definition of a successful outcome that was employed. Previous research that has investigated the relationships between earlier performance and post-school destinations and outcomes has tended to use a one-dimensional definition of a 'successful' outcome, focusing on participation in tertiary education or employment. The definition of 'success' used in this study was expanded by including satisfaction with life, as well as whether young people were fully occupied with education, employment or a combination of these activities, providing a more well-rounded view of outcomes than has been used in the past.

11.4.1 Engagement in Education, Training and Employment

The main activities of those low and high performers who remained in the study in 2007 were classified as being representative of full engagement (full-time work – 35 h or more on average per week; full-time study or training; part-time students who were working part-time or full-time hours), partial engagement (those working less than 35 h per week on average, part-time students who were not employed) or non-engagement (those who were looking for work but not employed and those who were not looking for work but not employed – not in the labour force). The proportions of the low performers who fell into these groups are presented in Table 11.1.

Overall, the outcomes in terms of engagement in education or employment for this group of young people appear fairly positive, with around seven in ten fully engaged in education or training, employment or a combination of these. However, in comparison to estimates for the full Y03 cohort and published statistics for the population of comparable age, the situation for this particular group of young people begins to look less favourable.

In 2007, 83% of the full Y03 cohort were fully engaged in education, training and/or employment, while 12% were partially engaged in these activities. Only 5% of the full Y03 cohort were not engaged in education, training or employment, half the proportion of the low-performing sample who were not engaged in these activities

[2] As this group of young people were already a sub-group of the original LSAY sample, use of the existing sample or attrition weights for the Y03 cohort was inappropriate.

Table 11.1 Level of engagement in employment, education and training of low performers, by background variables

Activity status in 2007 (age 19)	Fully engaged		Partially engaged		Unengaged	
	n	%	*n*	%	*n*	%
Gender						
Male	598	78.6	106	13.9	57	7.5
Female	586	67.0	183	20.9	106	12.4
Indigenous status						
non-Indigenous	1,100	73.1	264	18.9	140	9.3
Indigenous	84	63.6	25	18.9	23	17.4
Year 12 certificate						
No	393	69.4	91	16.1	82	14.5
Yes	791	73.9	198	18.5	81	7.6
School location						
Metropolitan	816	72.2	211	18.7	103	9.1
non-Metropolitan	368	72.7	78	15.4	60	11.9
Total	1,184	72.4	289	17.7	163	10.0

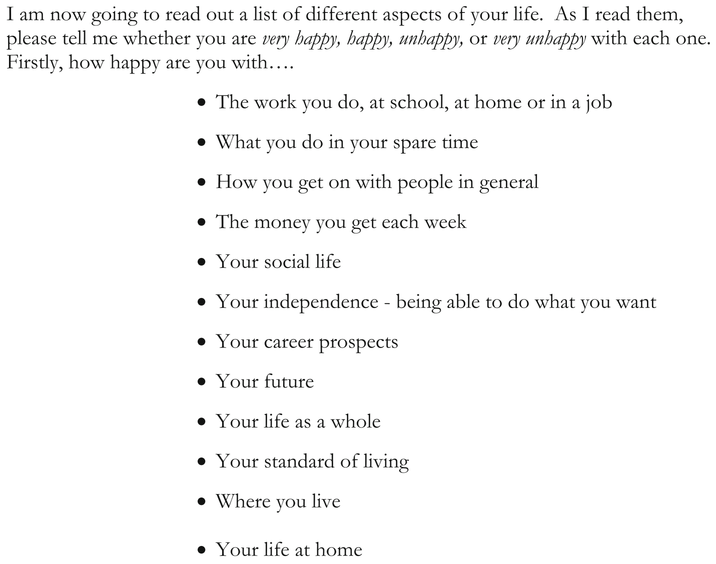
in that year (see Table 11.1). Australian Social Trends 2005 (ABS, 2005) reported on the engagement of different groups of young Australians and found that only 14% of young people aged between 15 and 19 were not fully engaged in 2004, which rose to 31% when only those who had left school in the previous year were considered. In comparison, almost 30% of this group of young people was not fully engaged in 2007, although the majority had actually left school late in 2005.

11.4.2 Successful or Not – Investigating the Differences

As mentioned earlier, the definition of success used in this study involved young people being fully engaged, as defined in the previous section, as well as being happy with their lives. Each year they are interviewed, LSAY participants are asked a series of questions about how happy they are with various aspects of their lives. Despite the use of the term ‘happy’, this measure corresponds more closely with the cognitive aspect of emotional well-being (life satisfaction) than with the affective aspect of emotional well-being (happiness). These items are presented in Fig. 11.1.

Responses to these items were coded (4 for very happy, 3 for happy, 2 for unhappy and 1 for unhappy) and the average response across the 2007 items calculated for each individual in the sample.[3] This score was then compared to

[3] Previous research with the LSAY data that has used these variables has reported that all of the items load together sufficiently in factor analyses (Hillman & McMillan, 2005) as to be used in this way.



I am now going to read out a list of different aspects of your life. As I read them, please tell me whether you are *very happy, happy, unhappy,* or *very unhappy* with each one. Firstly, how happy are you with….

- The work you do, at school, at home or in a job
- What you do in your spare time
- How you get on with people in general
- The money you get each week
- Your social life
- Your independence - being able to do what you want
- Your career prospects
- Your future
- Your life as a whole
- Your standard of living
- Where you live
- Your life at home

Fig. 11.1 Questions asked in the LSAY surveys regarding life satisfaction/happiness

the average response for the entire Y03 cohort in 2007 (the mean for the entire cohort was 3.42) and those members of the low-performing sample who scored at or above this level (equivalent to a response between 'happy' and 'very happy' across all items) were classified as happy for the purposes of the outcome variable.

Previous research with older LSAY cohorts has found an association between levels of engagement in activities and life satisfaction, with higher levels of satisfaction reported by those young people who are fully engaged with education, training or employment or some combination of these activities compared to young people who are only partially engaged or not engaged in such activities (see Hillman & McMillan, 2005). Among the young people in this analysis, there was an association between full engagement and being happy, with higher proportions of those who were fully engaged also meeting the criteria for being happy, particularly in comparison to those who were not engaged in any education, training or employment activities when interviewed in 2007.

Those young people who were fully engaged in 2007 (full time work; full time study or training; part time students who were working part time or full time hours), and whose responses to 12 items presented in the annual survey measuring satisfaction with various aspects of their personal lives indicated that they were happier than average (compared to the entire cohort in 2007) were designated as having a 'successful outcome' and thus formed the samples for the subsequent analyses – 1,596 low-performing students from 294 schools and 1,714 high-performing students from 288 schools.

11.4.3 Variables and Analysis

Multilevel logistic analysis was used to examine what factors differentiated between the 602 low-performing sample members who had a successful outcome (in terms of their level of engagement and happiness) and those sample members with less positive outcomes. The sample for this analysis included 1,596 students from 294 schools.

Similarly, multilevel logistic regression analysis was used to examine the factors which differentiated successful outcomes for those students who achieved Proficiency Level 5 or Proficiency Level 6 in mathematics. The sample for the analysis of high achievers included 1,714 students from 288 schools.

The following student (Level 1) characteristics were tested in the modelling.[4] The source of the item is indicated. For all categorical or dichotomous variables the first category is considered the reference group.

11.4.3.1 Young People's Background Variables

- Gender (PISA: female, male)
- Indigenous (PISA: no, yes)
- Have Year 12 certificate (LSAY: no, yes)
- Socioeconomic background (PISA). This analysis used the index of economic, social and cultural status (ESCS), which was created in PISA to capture the wider aspects of a student's family and home background. The ESCS is based on the highest level of the father's and mother's occupations, the highest level of education of the father and mother converted into years of schooling; the number of books in the home; and access to home educational and cultural resources. This was divided into quartiles based on data for the whole cohort and then two dummy variables were created: medium SES (which combined the middle two quartiles) and high SES, meaning low SES was the reference group used.

11.4.3.2 Student Motivation Variables

Two indices were developed in PISA to assess students' motivation to learn mathematics. The interest in mathematics index focuses on students' own, or internal, motivations to learn, and the instrumental motivation in mathematics index focuses on the external rewards that encourage students to learn. These indices were scaled

[4] It is acknowledged that not all of the young people could accurately be described as 'students' in 2007, however because the bulk of the variables included at this level of the model were indeed collected while the young people were students, this is the term that will be used to describe Level 1 influences.

using a weighted maximum likelihood estimate (OECD, 2004). Values on the index were standardised so that the mean value for the OECD student population was zero and the standard deviation was one. Thus negative responses on these indices indicate a response that was more negative than the OECD average.

Interest in mathematics (PISA). In this set of items students were asked to think about their views on mathematics and indicate their agreement on the following statements:

- I enjoy reading about mathematics.
- I look forward to my mathematics lessons.
- I do mathematics because I enjoy it.
- I am interested in the things I learn in mathematics.

Instrumental motivation (PISA). Students' levels of instrumental motivation were measured by seeking their responses to statements about the importance of mathematics for their future study and career prospects. Students were asked their level of agreement for each of the following statements:

- Making an effort in mathematics is worth it because it will help me in the work that I want to do later on.
- Learning mathematics is important because it will help me with the subjects that I want to study further on in school.
- Mathematics is an important subject for me because I need it for what I want to study later on.
- I will learn many things in mathematics that will help me get a job.

Two other variables used in the analyses were part of the LSAY questionnaire and broadly examined students' perceptions of the quality of school life. The items were Likert scaled and the score for the construct was formed as the average of the items that comprised the scale. The scales were:

Positive Affect – Your school is a place where:

- you feel happy;
- you like learning;
- you get enjoyment from being there;
- you really like to go each day;
- you find that learning is a lot of fun;
- you feel safe and secure.

Opportunity: Your school is a place where:

- the things you learn are important to you;
- the work you do is good preparation for your future;
- you have gained skills that will be of use to you;
- the things you learn will help you in your adult life;
- you are given the chance to do work that really interest you;
- the things you are taught are worthwhile.

11.4.3.3 Perceived Classroom Climate Variables

Two variables from the LSAY questionnaire were used to examine the effect of students' perceptions about the level of orderliness in the classroom, and the quality of teaching and of teacher-student relationships.

Student behaviour (LSAY). This variable was the average response to four items: Your school is a place where students

- are eager to learn;
- work hard;
- make good progress; and
- are well behaved.

Teacher-student relationship (LSAY). This variable was the average response to six items: Your school is a place where teachers

- know their subject matter well;
- explain things clearly;
- are well prepared and organised;
- have ability to communicate with students;
- maintain student interest; and
- manage student discipline well.

11.4.3.4 Aspiration Variable

Post-school plans (LSAY). In their initial LSAY survey, young people were asked about their plans for the future. Four dummy variables were developed, including the reference group who planned attending university. The other groups were:

- plan to do apprenticeship or traineeship,
- plan to go on to Technical and Further Education (TAFE),[5]
- plan to get a job, and
- don't know.

11.4.3.5 School-Level Variables

At the school level, four variables were used in the modelling. These variables together provide a contextual background for students in terms of school climate:

[5] In Australia, Technical and Further Education (TAFE) institutions are a government-owned system of colleges that offer post-secondary qualifications, often with a practical training focus (e.g. vocational education and training). Courses are developed in consultation with business and industry.

where their school is located, the type of neighbourhood and two measures of classroom climate – the general feelings about student behaviour and teacher-student relations at the school (among 15-year-olds).

- School location (PISA: Metropolitan, non-Metropolitan)
- School-average socioeconomic background. This variable was aggregated from the student-level socioeconomic background for the cohort.
- School-level student behaviour. This variable was aggregated from the student-level responses to these items for the subsample.
- School-level teacher-student relationships. This variable was aggregated from the student-level responses to these items for the subsample.

11.5 Results

Figure 11.2 shows the results for the low-performing group graphically.[6] In this figure, the solid bars represent the odds ratio of the event, and the lines represent the confidence interval around this odds ratio. Statistically significant odds ratios are indicated with an asterisk. In this section we will refer to both the calculated odds ratios and the associated predicted probabilities. For the reference group an odds ratio of 1 and the associated predicted probability[7] of 0.5 means that success is as likely as failure, thus odds ratios significantly higher (or lower) than 1, with associated predicted probabilities higher (or lower) than 0.5, mean that success or failure are more (or less) likely.

Of the background variables, only socioeconomic background was found to have a statistically significant association with success. Low-performing young people from medium and high socioeconomic backgrounds were more likely to be successful than young people from a low socioeconomic background. For those from an average socioeconomic background, the odds ratio was 1.3. Gender and Indigenous status were not found to be significant correlates of the likelihood of success among low-performing youth, and neither was the attainment of a Year 12 certificate.

Of the student motivation variables, two were found to be significant influences on outcomes: Positive Affect, the extent to which students reported enjoying being at school and learning, and Instrumental Motivation, or how important students thought mathematics would be for their future. The predicted probability of a successful outcome for young people in the low-performing sample with a higher score on Positive affect was 0.59 and for those with a higher score on Instrumental motivation, 0.54.

[6] The data behind these graphs are provided in the Appendix.

[7] The predicted probability is calculated as probability = odds/(1 + odds).

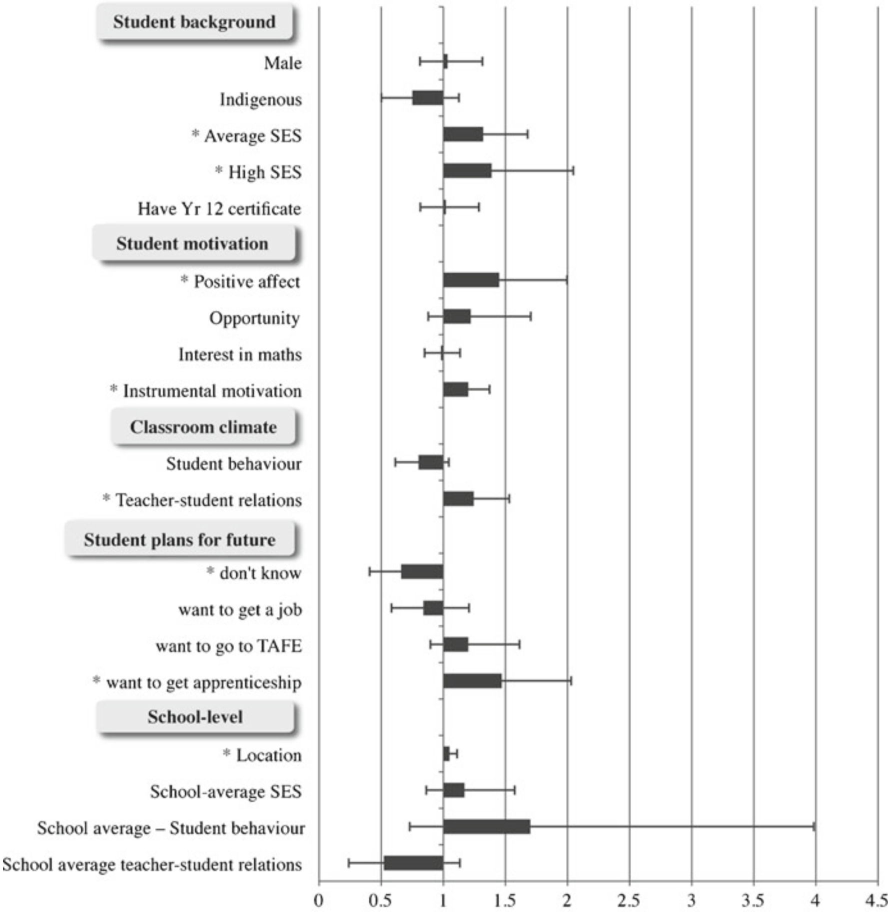


Fig. 11.2 Odds ratios for multilevel model of low performers' successful outcomes

Of the perceived classroom climate variables at the student level, only perceived teacher-student relationships were found to be significant, with those young people perceiving a more positive classroom climate more likely to be successful in later years.

In terms of aspirations or plans for the future, the expressed aim of obtaining an apprenticeship was associated strongly and positively with later success, while not having any definite aim was found to be significantly negatively related to success, with the probability of success for those young people answering "I don't know" to this question around 0.4.

Finally, of the school level variables investigated, the only one that was found to have a significant influence was location. Young people who had attended schools in a non-metropolitan location were found to be significantly more likely to be successful than those who had attended schools from a metropolitan location, other things equal. While Student-Teacher relationships was a significant influence on outcomes at the student level, it was not significant at the school level.

The significant role these variables had in the outcomes for these young people was explored further by replicating the analyses with a sample of high-achieving youth from the same cohort. This high-performing sample consisted of 1,714 students who scored at or above Proficiency Level 5 on the PISA 2003 mathematics test, from 288 schools.

The results of the logistic regression analysis are shown graphically in Fig. 11.3. None of the student background variables had an influence on the likelihood of success for this group of high-performing young people, nor did their plans for the future – almost 80% of the high achievers planned to go on to university and this lack of variation in post-school aspirations is probably behind the finding of no influence. Among the student motivation variables, Positive Affect and Opportunity both had a positive influence on the likelihood of these young people being ‘successful’ in the years after leaving school, but there was no significance influence of Instrumental Motivation, a direct contrast to the findings for the low-performing students. Positive relationships between students and teachers had a significant influence at the individual level and at the school level, indicating that high-performing students benefit not only from their own positive relationships with their teachers but also from being in a school in which other students get along well with staff.

Table 11.2 summarises the findings of the models for low and high performers. Those variables that had a positive influence on the outcomes for lower performing students that did not appear to make a difference to outcomes of high-performing students included socioeconomic background (low performers from average or high socioeconomic background had a greater likelihood of success); post-school plans, particularly aspiring to undertake an apprenticeship or traineeship for males, or simply having a plan formulated for females (as opposed to not knowing what they wanted to do); and instrumental motivation, seeing the value that mathematics has for everyday life and for their future plans.

11.6 Discussion

For all students:

The results of these two analyses highlight the importance a positive and supportive school climate, in which all students, low and high performers alike, can find a teacher with whom they can get along, can have on outcomes in their future lives. Ensuring that the school experience is a positive one not only impacts on young people’s lives at the time they are at school but appears to continue to influence them once they have left. While it is not possible to eliminate all stress or negative experiences

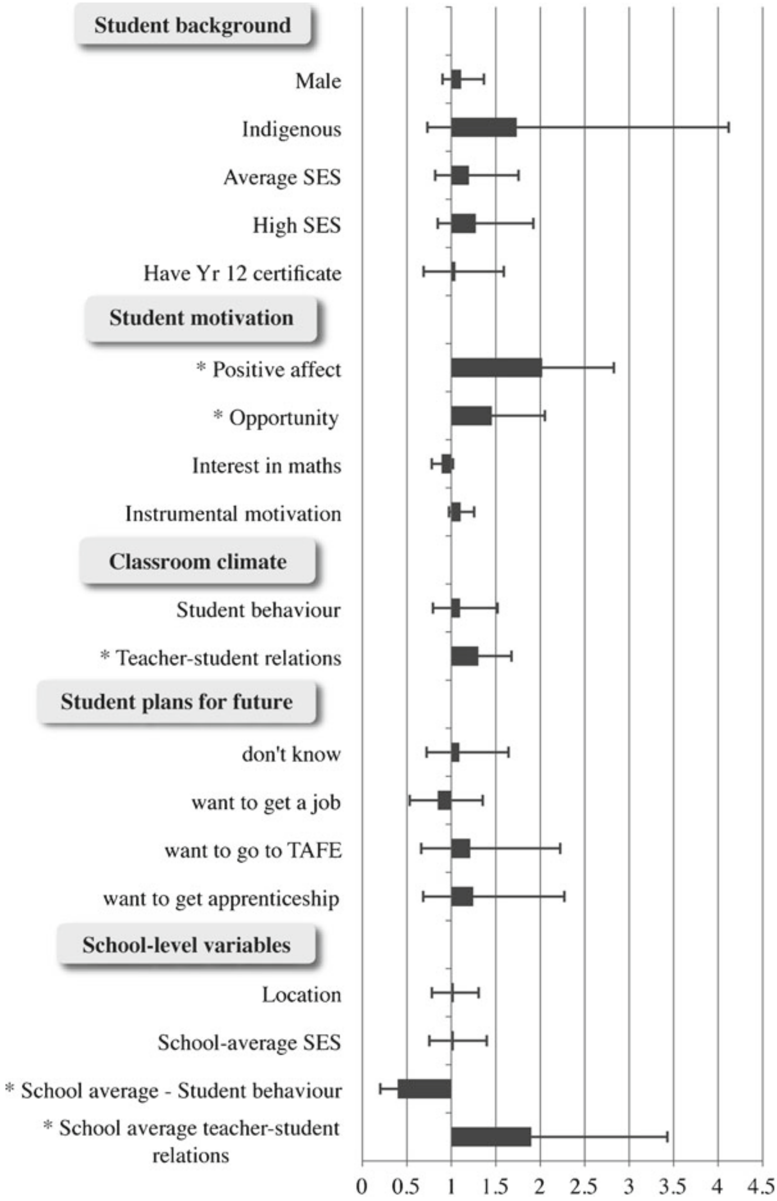


Fig. 11.3 Odds ratios for multilevel model of high performers' successful outcomes

from secondary school, findings such as this remind us of the important aim of education to foster the social and emotional development of young people, as well as their academic development, and that school can be a positive experience for all students, regardless of their achievement level, if the appropriate balance is found between

Table 11.2 Significant influences on successful outcomes for low and high mathematics performers

Significant influence on success	Low performers	High performers
Student background		
	Average socioeconomic status (positive)	–
	High socioeconomic status (positive)	–
Student motivation		
	Positive affect (positive)	Positive affect (positive)
	–	Opportunity (positive)
	Instrumental motivation (positive)	–
Classroom climate		
	Teacher-student relations (positive)	Teacher-student relations (positive)
Plans for future		
	Do not know (negative)	–
	Want to get apprenticeships (positive)	–
School-level		
	Non-metropolitan location (positive)	–
		School average student behaviour (negative)
		School average teacher-student relations (positive)

encouraging the pursuit of personal goals and development, and comparison and ranking of student achievement. Khoo and Ainley (2005) have reported also that positive attitudes towards school and plans to continue with education have a positive effect on actual participation in further study, above the influence of earlier achievement on participation.

For low performers:

At the same time, young people who may not be performing as well as their peers should be encouraged to think carefully about their future and to make strategic plans. Those young people, particularly females, who were not performing well in mathematics and who had not thought about what they might do after leaving school were much less likely to be fully engaged and happy with their lives 4 years down the track.

The significant influence of instrumental motivation on the group of low-performing students' later outcomes is an important message for parents, teachers and policy-makers. Finding that lower performing students who recognise the value of mathematics for their future success are more likely to achieve this success, and that includes being happy with many aspects of their personal lives as well as their future and career, suggests that a focus on the practical applications of mathematics in everyday life may go some way to improving the outlook for students who are not quantitatively inclined and are not achieving well in the mathematics classroom. Other research (e.g. Samuelsson, 2008) has illustrated how different teaching strategies can be used to increase students' levels of instrumental motivation as well as their performance in various mathematics tasks, indicating that it is indeed possible for educators to have a positive effect on something that may well improve their students' future outcomes outside of the mathematics classroom.

Appendix: Odds-Ratio Coefficients for Logistic Regression Analyses

Table A.11.1 Multilevel odds-ratio coefficients and confidence intervals – low-achieving students

	Odds ratio	Confidence interval	
Student background			
Male	1.0	0.81	1.32
Indigenous	0.8	0.50	1.12
*Average SES	1.3	1.04	1.68
*High SES	1.4	1.00	2.05
Have Year 12 certificate	1.0	0.81	1.29
Student motivation			
*Positive affect	1.4	1.05	2.00
Opportunity	1.2	0.88	1.70
Interest in Maths	1.0	0.85	1.14
*Instrumental motivation	1.2	1.05	1.37
Classroom climate			
Student behaviour	0.8	0.61	1.04
*Teacher-student relations	1.2	1.01	1.53
Student plans for future			
*Don't know	0.7	0.41	0.99
Want to get a job	0.8	0.58	1.21
Want to go to TAFE	1.2	0.90	1.61
*Want to get apprenticeship	1.5	1.07	2.03
School-level variables			
*Location	1.1	1.01	1.11
School-average SES	1.2	0.86	1.57
School average – student behaviour	1.7	0.73	3.98
School average teacher-student relations	0.5	0.24	1.13

*p<0.05

Table A.11.2 Multilevel odds-ratio coefficients and confidence intervals – low-achieving students

	Odds ratio	Confidence interval	
Student background			
Male	1.1	0.90	1.37
Indigenous	1.7	0.73	4.12
Average SES	1.2	0.82	1.76
High SES	1.3	0.85	1.92
Have Year 12 certificate	1.0	0.69	1.59
Student motivation			
*Positive affect	2.0	1.45	2.83
*Opportunity	1.5	1.03	2.05
Interest in maths	0.9	0.78	1.02
Instrumental motivation	1.1	0.97	1.26
Classroom climate			
Student behaviour	1.1	0.79	1.52
*Teacher-student relations	1.3	1.02	1.68
Student plans for future			
Don't know	1.1	0.72	1.64
Want to get a job	0.8	0.53	1.35
Want to go to TAFE	1.2	0.66	2.22
Want to get apprenticeship	1.2	0.68	2.27
School-level variables			
Location	1.0	0.78	1.31
School-average SES	1.0	0.75	1.40
*School average – student behaviour	0.4	0.20	0.78
*School average teacher-student relations	1.9	1.05	3.43

* $p<0.05$

References

Australian Bureau of Statistics. (2005). *Australian social trends, 4102.0 Education and work: Young people at risk in the transition from school to work*, http://www.abs.gov.au/Ausstats/abs@.nsf/0/18173651BD0E7F4FCA25703B0080CCC2?Open. Accessed 27 Mar 2009.

Curtis, D., & McMillan, J. (2008). *School non-completers: Profiles and initial destinations* (LSAY Research Report 54). Melbourne, Australia: Australian Council for Educational Research.

Fullarton, S. (2002). Student engagement with school: individual and school-level influences, (LSAY Research Report 27). Melbourne, Australia: Australian Council of Educational Research.

Fullarton, S., Walker, M., Ainley, J., & Hillman, K. (2003). *Patterns of participation in Year 12* (LSAY Research Report 33). Melbourne, Australia: Australian Council for Educational Research.

Hillman, K., & McMillan, J. (2005). *Life satisfaction of young Australians: Relationships between further education, training and employment and general and career satisfaction* (LSAY Research Report 43). Melbourne, Australia: Australian Council for Educational Research.

Khoo, S. T., & Ainley, J. A. (2005). *Attitudes, intentions and participation* (LSAY Research Report 41). Melbourne, Australia: Australian Council for Educational Research.

MCEETYA Performance Measurement and Reporting Taskforce (PMRT). (2006). *Measurement framework for national key performance measures*. Available at www.mceetya.edu.au/verve/_resources/2006_Measurement_FW_for_national_KPMs_Final.pdf. Accessed Apr 2009.

McMillan, J., & Marks, G. N. (2003). *School leavers in Australia: Profiles and pathways* (LSAY Research Report 31). Melbourne, Australia: Australian Council for Educational Research.

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

Samuelsson, J. (2008). The impact of different teaching methods on students' arithmetic and self-regulated learning skills. *Educational Psychology in Practice, 24*(3), 237–250.

Thiessen, V. (2007). *The impact of factors on trajectories that lead to a high school diploma and to participation in post-secondary education among those with low reading competencies at age 15* (Report for Learning Policy Directorate, Strategic Policy and Research, Human Resources and Social Development Canada). Available at http://www.hrsdc.gc.ca/eng/publications_resources/learning_policy/sp_795_11_07e/page00.shtml. Accessed Mar 2009.

Chapter 12
Linking PISA Competencies over Three Cycles – Results from Germany

Claus H. Carstensen

Abstract Since the publication of the PISA 2006 study results the question of reporting trends over the PISA cycles has received a lot of interest. This chapter discusses the possibilities and limitations of trend analyses based on data from this international comparative study and using complex test designs. The chapter succeeds trend analyses which were carried out with the German data from the first three PISA studies in 2000, 2003 and 2006 (Carstensen CH, Prenzel M, Baumert J, Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt? [Trend analyses in PISA: how did competencies in Germany develop between PISA 2000 and PISA 2006?] Zeitschrift für Erziehungswissenschaften, Sonderheft 10:11–34, 2008; Prenzel M, Artelt C, Baumert J, Blum W, Hammann M, Klieme E et al (eds), PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie [PISA 2006. Results of the third international comparison]. Waxmann, Münster, 2007).

The choice of a scaling and trend analysis model depends on the focus of the analysis and on the assessment design. With respect to international comparisons, very strict assumptions on the uni-dimensionality of the test instruments used have to be made to allow for trend analyses. What if these conditions are not met across all participating countries for all assessment cycles? This paper presents an alternative model for trend analyses, assuming uni-dimensionality only within a particular country but not across all participating countries. Trend results with this model can only be interpreted within the particular country and are not intended for use in international comparisons.

C.H. Carstensen (✉)
Institute of Psychology, University of Bamberg, Bamberg, Germany
e-mail: claus.carstensen@uni-bamberg.de

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_12,

To establish the validity of the presented trend model, an empirical analysis of the different tests and subscales used in different assessment cycles was performed. As far as different versions of the instruments were administered within cycles, the correlations of these test forms give an empirical indication of the uni-dimensionality of the underlying constructs. Monte Carlo simulations were performed to analyse whether the correlations of these test forms indicate a uni-dimensional construct being measured over time. Having analyzed the correlations of the tests, a fit analysis at the item level followed. Further assumptions refer to the stability of item difficulties over time. This is addressed by estimating item by time interaction parameters, allowing for a descriptive analysis of items changing their difficulty over time and a model fit comparison to check whether item drift has an impact on their difficulties.

Results show that trends might be reported for the German data, using the short test for reading and using all pair wise link items in Mathematics and Science. In the conclusion, the results and some implications for the design of future PISA assessments will be discussed.

Keywords Trends • Scaling models • Measurement invariance

12.1 Issue: Trends from PISA Data

From a methodological perspective the question of trends is clearly distinct from the question PISA data have to answer in the first place: each PISA cycle compares student performance with the purpose of country comparisons. For country comparisons within a PISA cycle, the first optimization criterion for study design, scaling and analysis procedures would be the comparability across countries. The competencies measured need to have the same meaning and be interpreted identically within each country. In contrast, for trend analyses, the highest priority in optimizing study design, scaling and analysis procedures would be comparability across cycles; the competencies found in each cycle need to have the same meaning and be interpreted identically across cycles. In case researchers find that the proficiency distributions from a trend study cannot be analyzed under both perspectives, they will have to decide according to which perspective the data analysis procedures shall be optimized.

The rationale of trend analyses is to keep the instrument and assessment conditions the same across different studies and then to assume that any change in the item response frequencies is due to a change of proficiencies of the sampled populations. Hence, a prerequisite of trend analyses is to prove the equal characteristics of the common instrument or measurement invariance across studies (Kolen & Brennan, 2004). In the remainder of this section the invariance of the measurements within the first three PISA cycles will be discussed.

In order to provide differential information on student's competencies, the focus of the competence assessments varies across the cycles: in PISA 2000, reading literacy was the major domain of the study, which can be seen from the number of items among other criteria. In PISA 2003, mathematical literacy was the major

Table 12.1 Domains, numbers of items and framework development for the first three PISA cycles

	PISA 2000	PISA 2003	PISA 2006
Reading	**Major** 129 items full framework	Minor 28 link items 00/03/06	Minor 28 link items 00/03/06
Mathematics	Minor 20 link items 00/03	**Major** 84 items full framework	Minor 48 link items 03/06
Science	Minor 22 link items 00/03	Minor 25 link items 03/06	**Major** 108 items full framework

domain, and scientific literacy in PISA 2006. Thus, different test instruments were used in different cycles: instruments of the major domain comprise a large number of items; these tests are called long tests in this chapter. If a domain was assessed not as the major domain, a smaller number of items was administered in the assessment; these tests are called short tests in this chapter.

With each of the three domains being a major domain over the course of the 2000, 2003 and 2006 PISA cycles, the fully elaborated assessment frameworks were worked out successively in parallel with the study focus: the fully detailed framework for mathematical literacy was presented 2003 (OECD, 2003) and the fully detailed framework for science was presented in 2006 (OECD, 2006). As a consequence, the short tests in Mathematics and Science administered in the studies before these domains were the major domain are not necessarily subsets of items of the respective long tests. Moreover, the selection of items for the short tests could not be related to the respective long tests and the number of common items between cycles is smaller than necessary for a stable link. In contrast, the short test in reading was designed based on the results of the long test.

Table 12.1 gives an overview over the three domains and the three cycles, as well as over the framework development, the number of items in the long tests, and the number of common items.

The same short test for Reading was administered in PISA 2003 and PISA 2006. It consists of 28 items selected from the PISA 2000 reading assessment. For mathematics, 84 items were administered in the long test and 48 of these items were selected for the PISA 2006 assessment. However, only 20 items of the 34 items of the PISA 2000 Mathematics assessment also appear in the long test. Only eight items appear in both the PISA 2000 and the PISA 2006 Mathematics assessments. The long test in Science includes 108 items, the PISA 2003 short test consists of 25 items which also appear in the PISA 2006 assessment and nine unique items. The PISA 2000 science assessment has 22 items in common with PISA 2003 and 12 items in common with PISA 2006 and no unique items.

Given the assessment design of the three PISA studies, the OECD (2007) reported trends over the first three cycles for reading only, trends for Mathematics were reported from PISA 2003 to PISA 2006 and for Science no trends were reported. Gebhardt and Adams (2007) investigated the impact of different scaling and linking methodology on trend results. They found that using different instruments in different

Table 12.2 Mean difficulties for the common and unique items from PISA 2000

Reading assessment mean difficulties	OECD	Germany	Sweden	Mexico
Link items 2000/2003	–0.03	–0.05	–0.20	0.18
Unique items 2000	0.01	0.01	0.06	–0.05
Relative difficulty link items	0.04	0.06	0.26	–0.23

cycles for the same competence, like the long reading test and the short test in PISA 2003 and PISA 2006 may have led to biased results.

Gebhardt and Adams (2007) investigated item difficulties across countries for the first two assessment cycles (PISA 2000, 2003). They compared the mean difficulty of the common and the unique parts of the reading and science assessments across countries. Within all OECD countries, the difference between the common and the unique items in Reading in PISA 2000 is 0.04 logits, so the common items are slightly easier than the unique items. In the PISA 2003 reading assessment, only the common items were administered again. Therefore, students participating in PISA 2003 have a slight advantage in solving the items compared to the PISA 2000 participants. Through appropriate scaling and linking (OECD, 2005, 2009), however, the scale scores are in the same metric. If we look at a particular country, though, the difference between the common and the unique items appears different. For example, the mean item difficulties for Germany, Sweden and Mexico are presented in Table 12.2. In Sweden, the common items presented in PISA 2003 are 0.26 logits easier than the unique items, which in turn is 0.21 easier than in the OECD. Consequently, the Swedish students gain an advantage from switching to the short test. For Mexico, the short test is harder than the long test by 0.23 logits and consequently the disadvantage of the Mexican students in PISA 2003 is 0.27 logits (Gebhardt & Adams, 2007). For Germany, the difference to the OECD is only 0.02 logits and hence there is hardly any advantage or disadvantage gained from switching to the short test.

Gebhardt and Adams investigated the impact of these advantages and disadvantages on the trend estimates. They compared trend results from three different methods: the original scaling reported in OECD publications and two further scaling methods. Both latter methods (which will be further illustrated below) include the rescaling of the data for each country, so that the mean difference between the common and the unique items is modeled for each country individually and does not reflect the average OECD value of this difference. The main results of their study are that trend results are significantly different between the original scaling (with reference to the OECD value of the mean difference) and the alternative methods with country-specific mean difference treatment in 6 out of 28 countries. For the science assessments from PISA 2000 and PISA 2003, Gebhardt and Adams found significant differences in trend results for 2 out of 25 countries. These results may indicate the extent to which trend results in the PISA studies are variable conditionally on the scaling method.

In addition, other factors might have an impact on trend results as well. Gebhardt and Adams investigated the influence of different sample characteristics (such as the number of public and private schools or the distribution over socio-economic backgrounds) between cycles on trend results. Another factor might be seen in item-by-study-by-country interactions in the item difficulties, like item drift over time. Further assumptions for comparability across cycles have to be made with respect to the booklet design, i.e. the rotation of clusters within booklets and the positions of items within a cluster have to be the same for the common items.

As said before, trend analyses of the data collected in the PISA cycles require measurement invariance. Measurement invariance in IRT models can be assumed if the same item response model holds for all measurement occasions, i.e. for all studies in all countries, which can be assumed if the item difficulty parameters are the same for each item across studies and across countries. As shown by Gebhardt and Adams, this assumption does not hold with respect to the mean item difficulties between common and specific items for all countries. This chapter will investigate whether a model for trend analyses without assuming item parameter equality within and across cycles will allow trend analyses within a country. The research question in this chapter is whether an appropriate scaling method (modeling the mean differences mentioned above for each country) will prove to be a reliable basis for trend analyses for the German PISA data across PISA 2000, PISA 2003 and PISA 2006.

12.2 A Model for National Trend Analyses

12.2.1 IRT Scaling Model

The trend model to be investigated in this chapter applies a concurrent scaling to the data of the three PISA cycles from German students only. This model was introduced as the marginal trends model by Gebhardt and Adams (2007). In the marginal trends model, the response data from the three PISA cycles of interest, that is common and unique items, are calibrated concurrently, with the item difficulties of common items being assumed to be equal across cycles.

The proficiency distributions were estimated in two steps. In a first model (model 1) item parameters were estimated from a dataset with student responses from three cycles for each domain using a Rasch type model for dichotomous responses and a partial credit model (Masters, 1982) for items with three- or four-point scores using the ConQuest 2.0 software (Wu, Adams, Wilson, & Haldane, 2007). For PISA 2000 data, booklet effects were estimated as well, since in PISA 2000 item difficulties are confounded with booklets (Adams & Carstensen, 2002). For PISA 2003 and PISA 2006, no booklet effects were estimated, since in these studies they have no impact on item parameter estimates and proficiency distributions (OECD, 2005, S. 198, 2007).

In a second step (model 2), item parameters were kept fixed and student proficiency distributions were estimated conditionally on study, type of school and their interaction effects. Plausible values (Adams & Wu, 2002) were derived and transformed into a metric for reporting trends. A metric with a mean of 100 and a standard deviation of 30 was chosen to make it obvious that these estimates were not obtained from the OECD scaling model.

This trend model differs from the model of Gebhardt and Adams (2007) in two respects: their conditioning model includes the students' age and gender, the socio-economic status of their parents, migration background and dummy variables for missing responses. Secondly, in this chapter three uni-dimensional models were used for reading, mathematics and science instead of the three-dimensional model of Gebhardt and Adams. Since trends will only be reported in terms of means and variances, no substantial differences are expected due to these minor differences between the trend models.

The proposed model analyzes data from all cycles concurrently. If results from earlier cycles have already been published, it might be impossible to report numerically different results for that wave from a new calibration. Xu and von Davier (2008) discuss different models for item parameter estimation in a comparable setup. If item parameters are fixed to their values from the first assessment, the linked item parameters for later cycles as well as the ability distributions and their changes differ significantly from the respective values if item parameters are estimated for all assessments concurrently. To extrapolate the trends proposed herein for the time after PISA 2006, one has to decide whether further concurrent calibrations may be performed with the data collected from further cycles, which may change results already published from previous cycles or if other linking models have to be adopted for continuing trend reports.

12.2.2 Trend Model Validity and Fit

The following section discusses whether crucial assumptions made in the trend model described above hold for the German data. This includes a discussion of the compatibility of the assessment frameworks across studies, an analysis investigating whether the common and unique items form uni-dimensional scales within each cycle and an analysis of item fit (both questions of construct validity) and an analysis whether item-by-study interactions are negligible (a question of trend model fit).

With the trend model presented, a uni-dimensional scale based on items across studies for each domain within a single country will be established. Note that with the cross-sectional comparisons reported, the assumption of uni-dimensionality has been assessed within cycles across countries. To validate a trend model, the definition of the domains across cycles needs to be consistent. Since the assessment frameworks have been developed over the studies, this consistency will be discussed for each domain. Reading was the focus in PISA 2000. The assessment instrument comprises 129 items and can be analyzed with respect to different reading subscales. In the 2003 and 2006 PISA assessments, the same short test was administered, consisting of 28 items[1] which were selected to represent the reading scale as a whole. Thus, the reading proficiency scales of the three PISA studies might be linked using the short test from all three cycles. In order to use all items available, the long test from PISA 2000 might be linked with the short test from the following two cycles. Whether the short test and the long test measure the same construct empirically can be analyzed from the PISA 2000 data. Results of this analysis will be reported in the following section.

The mathematics assessments used three different tests, two different short forms and a long test in PISA 2003. The long test included 84 items and differentiated four subscales. The short test for PISA 2006 consisted of 48 items balanced over these subscales. However, the PISA 2000 short test consisted of 31 items and basically includes two of the subscales. It is not balanced with respect to the mathematical content areas defining the subscales. Furthermore, it shares 20 items with the long test and only 8 items with the PISA 2006 short test. The OECD reports on trends in mathematics refer to the two subscales included in the 2000 short test. In contrast, the analyses presented here link both short tests to the PISA 2003 assessment. Whether these tests measure the same construct of mathematical literacy will be investigated as an empirical question in the following section.

The full framework for scientific literacy was developed for the PISA 2006 study. Nevertheless, it is largely consistent with the prior frameworks (OECD, 2006, p. 25), and the science tests from the three studies are thus constructed rather consistently with respect to the combined science score. The 2000 and 2003 short tests share 25 items. The long test shares 14 items with the PISA 2000 short test and 22 items with the PISA 2003 short test. Just as for reading and mathematics, results from an empirical analysis of the factorial validity will be presented for science in the following section as well.

12.3 Results from German Data

For using the trend model presented, items from different tests are selected and analyzed together. Hence, strictly speaking, new tests are constructed. For the domains of mathematics and science it is furthermore assumed that these new combinations of items measure a common construct within each domain. In order to investigate whether these assumptions hold, the results of empirical analyses of the factorial validity of the new tests and of item fit analyses are presented. Moreover, an analysis of whether the items from the three assessments in each domain form a common scale using item fit statistics will be undertaken. Finally, the trend results from German data will be presented.

[1] Due to deletion of one reading item for the German data set, the short test in the following analysis includes 27 items.

Table 12.3 Estimated correlations and bootstrap results for five links in the trend model: data source, contrast, observed correlations, smallest correlation from r = 100 bootstrap samples and number of unique and link items in the booklet design

Data	Contrast	Observed correlations	Smallest corr. from bootstrap	No of items: (unique/link)
PISA 2000	Reading unique vs. link 00/03/06	0.926	0.942	101/27
PISA 2003	Mathematics unique vs. link 00/03	0.944	0.948	64/20
PISA 2003	Mathematics unique vs. link 03/06	0.970	0.970	36/48
PISA 2006	Science unique vs. link 03/06	0.955	0.934	81/22
PISA 2006	Science unique vs. link 00/03/06	0.960	0.908	89/14

12.3.1 Empirical Analysis of the Factorial Validity

To assess empirically whether the common and the unique items from the tests linked in the trend model form uni-dimensional scales, the following analyses were performed: latent correlations were computed for a two-dimensional model which contrasts the unique items and the common items for each test in the trend model. The correlations between the common items and the unique items have a sampling variance due to sampling of responses and due to measurement error, especially for links based on small numbers of common items. If both sets of items measure the same construct, those correlations should be maximal, that is not significantly different from r = 1. To obtain confidence intervals for these correlations, a bootstrap procedure was applied. The bootstrap model had a simplified set-up without creating a design for non-administered items from a multi-matrix design. The number of items for each bootstrap model was chosen to correspond to the number of unique and link items for each link evaluated. Particularly, the bootstrap procedure was based on the average number of link and unique items administered to students through all booklets for a particular domain. The sample size for each domain reflected the sample sizes of the respective PISA assessments as defined by their booklet design. The data sets were generated according to a two-dimensional Rasch model in which true values for the item parameter and ability distributions were generated for each replication, thus implementing a non-parametric set up using the "simulate" option of the Conquest software. Standard PISA analyses of correlations do not reflect dependencies of item responses due to unit design, so neither does the bootstrap design. However, not reflecting item dependencies and, possibly, fatigue effects and others in the bootstrap design might result in higher correlations and thus might suggest too liberal decisions in detecting non-equivalence of link and unique items.

In Table 12.3, the estimated latent correlations and the results of the bootstrap procedure are shown with respect to five correlations: one for the link in reading across all three cycles; for mathematics one for the link between the first and the second cycles and one for the link between PISA 2003 and PISA 2006; for science one for each link between PISA 2003 and PISA 2006 and one for all three cycles. All correlations were computed from data taken from the study where a domain was the major domain. In the model for reading the booklet coefficients were omitted since they would have had no impact on the dimensionality. The smallest correlation from 100 generated data sets for evaluating the correlation between both parts of the reading test of PISA 2000 was found to be 0.942. The observed latent correlation for reading from PISA 2000 data cannot be found within the range of the correlations from 100 replications. Hence, the observed correlation is statistically lower than $r=1$ and the two dimensions, the set of common items and the set of unique items, are not the same. The assumption that both parts of the reading test can be seen as parts of the same instrument does not hold for the German data.

For mathematics, the picture is different: the observed latent correlations from both links are quite close to the values of the bootstrap analysis and it can be concluded that with a probability of around $p=0.01$ the observed correlation is from within the range of observable correlations if the generated correlation is $r=1$. For the trend model, the common and the unique items in mathematics are used for both links. Both correlations for science are well inside the range of correlations from the bootstrap and thus it can be concluded that both parts of the science tests measure the same construct in the German data. These finding are not completely in line with the expectations from reviewing the frameworks; for reading, a good connection between both parts of the assessment had been expected. The links in mathematics between PISA 2000 and PISA2003 and the links in science between PISA 2003 and PISA 2006 were expected to be a bit weaker. As a consequence of these analyses, the trend model for Germany in reading will be computed based on the common items only, i.e. the short test from all three cycles. For the trend model in mathematics and science, all common and unique items administered in one or more cycles will be used.

Assuming that the tests for the trend model are uni-dimensional, the fit of single items into the scales can be assessed empirically. The PISA consortium (Adams & Wu, 2002) uses, among others, the “weighted mean square residual fit index” (Wright & Masters, 1982), which basically evaluates the discrimination of each item. Inspecting these values for the items of the trend model under investigation, only a few items show indications of misfit: 2 out of 27 reading items, 5 of 95 mathematics items and 3 out of 124 science items show significantly low fit values. The total percentage of significant fit values is 4% which is less than expected assuming a conventional 5% error probability. Hence, no items have been removed from the trend models because of item fit.

With a trend analysis, the difficulties of item responses are evaluated over time. If all items become easier by the same degree, the change may easily be attributed to a change in a population’s proficiency. If, however, items change differently in their difficulty, that is if there is an item by study interaction, the change cannot be attributed to a single dimension. In order to evaluate whether item by study interactions have an impact on the trend model, the following analysis was performed: based on

Table 12.4 Linking errors for three domains: domain, link, link error and number of link items

Linking errors			
		Link 2000 2003	Link 2003 2006
Reading	Error	4.73	4.67
	# link items	27	27
Mathematics	Error	4.43	2.17
	# link items	20	48
Science	Error	4.38	3.33
	# link items	24	22

Note: The link errors are in a metric with SD = 100 to be compared to OECD PISA values

the scaling (calibration) model for estimating the trends, two further models were estimated for each domain. For both models, the item parameters are held fixed at values from the calibration model and no conditioning model is specified. With the first model, the study is included as a conditioning variable for item difficulties to capture overall changes in the populations over time; this model does not assume an item by study interaction. With the second model, item by study parameters are introduced additionally. Comparing the fit of both models allows evaluation of whether item by study interactions have a significant impact on the item difficulties of the common items estimated from the German data.

Evaluating the item by study parameter estimates descriptively, one finds that a small number of these estimates are larger than 0.3 logits or smaller than −0.3 logits. For the science test, most of these estimates are even in the range from −0.2 to 0.2 logits. Linking errors were computed for the link between each of the two pairs of consecutive cycles and each domain. These link errors are displayed in Table 12.4, in a scale with SD = 100 to enable comparison with linking errors for PISA reported by the OECD. Linking errors for original trends in PISA (OECD, 2005, table 12.28) vary from 1.38 points (mathematics from 2003 to 2006) to 5.31 points (reading from 2000 to 2003) in their reporting scales. Monseur and Berezner (2007) compute linking errors using jack-knife techniques to reflect the item structure in units, item by country DIF and partial credit items. They report link errors for reading at the country level from about 6 to 12 points, for Germany they find an error of 9.54 points. The link errors of the linking model proposed here for the German data are in the same magnitude as the linking errors for the international trend model.

Table 12.5 displays results for model fit comparisons for reading, mathematics and science. It lists the model estimated, the number of students, the difference in parameters between both models for each domain, the deviance (−2ln likelihood) and the CAIC information criterion (Bozdogan, 1987). The CAIC indicates a better fit of the model with the smaller index value and is computed with respect to the difference in likelihood and number of parameters of the models compared given a sample size of the analyzed data set. However, it does not make assumptions about the distribution of the index values and does not provide a test for the significance of differences.

Table 12.5 Model fit results for three domains: model, number of model parameters, deviance (= –2log$L$) and CAIC, the sample size is N = 14,624 for each model

Model	# of par.	–2 ln$L$	CAIC
Reading			
Item + cycle	4	322055	322096
Item + cycle + cycle x item	4 + 54	321765	322359
Mathematics			
Item + cycle	4	259540	259581
Item + cycle + cycle x item	4 + 68	259269	260013
Science			
Item + cycle	4	281626	281667
Item + cycle + cycle x item	4 + 46	281442	281954

For each domain in the first model, four parameters are estimated – a study mean, two differences between study means and a variance – while item parameters are fixed to their values from the calibration. With each second model, a parameter for each common item and occasion is estimated. In reading, 27 items of the long test were administered in PISA 2003 and in PISA 2006, resulting in 54 item by time parameters. For mathematics, 68 item x study parameters and for science, 46 item x study parameters were estimated. For all three domains, the CAIC values of the item by study interaction model are bigger than the index value from the non interaction model. This indicates a better fit of all three non-interaction models. Given these results and the comparison of the link errors of the proposed model with PISA original scaling link errors, the item by study interaction parameters are assumed to be negligible with respect to measuring trends and all common items are linked by restricting their difficulty to be the same over cycles.

12.3.2 Trend Results

In the following section, the trend results for German data estimated using the proposed trend model (Carstensen, Prenzel, & Baumert, 2008) are reported. Due to the estimation of country-specific item difficulty parameters, the proficiency scales reflect curricular and cultural characteristics of the German educational system. Therefore, the scale values are not to be directly compared to international PISA scale values. To remind the reader of this, the trend scale values are reported in a metric with a mean of 100 points for the reference study and a pooled standard deviation of 30 over all three cycles.

In Table 12.6, the trend results are printed for all three domains. According to our trend model, the mathematical competencies of the 15-year-olds in Germany have increased over cycles. The PISA 2003 mean is set to 100; the PISA 2000 mean of 93 points is significantly lower with standard errors of 0.8 and 0.7 for both means.

Table 12.6 Trend results for reading, mathematics and science in Germany: Mean, SE and SD for each study and a linear trend estimate

	PISA 2000			PISA 2003			PISA 2006			Trend	
	mean	*SE*	SD	mean	*SE*	SD	mean	*SE*	SD	mean	*SE*
Mathematics	93	*0.8*	26	100	*0.7*	32	101	*1.2*	31	4.4	*0.7*
Science	100	*0.7*	26	104	*1.1*	33	107	*1.1*	30	2.4	*0.8*
Reading	100	*0.8*	26	100	*1.0*	32	100	*1.3*	32	–0.0	*0.7*

Notes: Means constrained on M = 100 for one assessment cycle, standard deviation fixed to SD = 30 over three cycles; concurrent calibration for German data

The PISA 2006 mean at 101 points is numerically higher than the PISA 2003 mean, but this difference is not statistically significantly different from zero. A linear trend was estimated over the three cycles (based on a dataset with plausible values from all three cycles) as well. The linear increase between two cycles is 4.4 points, which equals d = 0.15 in terms of effect size. Given its standard error of 0.7 points, this increase is statistically significantly larger than zero: on average, there is a significant increase in mathematical literacy over cycles in Germany.

The OECD reported trends in mathematics on the overall scale between PISA 2000 and PISA 2003[2]; in these reports, Germany gained about one point (in the SD = 100 metric; see Prenzel et al., 2007) which converts to 0.24 points in the SD = 30 metric. Comparing the two trend estimates, we find positive values neither of which are significantly different from zero.

For scientific literacy the German trend estimates show an increase over cycles: with the mean value for PISA 2000 being fixed to 100, the means increase to 104 and 107 points respectively, and the linear trend is a 2.4 points increase between cycles. All mean differences between cycles as well as the trend are significantly different from zero. Gebhardt and Adams (2007) also report a significant increase between PISA 2000 and PISA 2003 for science in Germany; in the OECD scaling (OECD, 2004) we find a significant increase between these two study means as well. As far as any are available, the different trend estimates consistently show an increase in scientific literacy.

The reading proficiency of the fifteen-year-olds in Germany did not change significantly over cycles. The mean value from PISA 2000 is set to 100, the mean values from the other two cycles are both 100 points as well, so the linear trend is also zero. This result is somewhat in contrast to the trend computed from the OECD scaling, in which the study means for Germany are 484, 491 and 495 points, showing a numerical increase. However, neither of the study mean differences are statistically significant. Gebhardt and Adams found no increase in reading literacy from PISA 2000 to PISA 2003 in Germany, which is consistent with the national trend

[2] Earlier trend reports were restricted to two subscales which were assessed with sufficiently large item numbers.

model presented in this chapter. This inconsistency between national trend estimates and trends from OECD scaling values is due to the different trend models implemented. The national model re-estimates item parameters specifically for a country, which is in case of differences the model fitting more closely and thus more reliable. Trends from OECD scaling values suffer from item by country interactions in item difficulty and especially from combining the use of different test forms (long and short form). Making the PISA test comparable across countries reduces the stability of trend analyses if the item difficulties vary over countries. Again, such variation might be due to cultural and school system factors and could be the result of sampling variation of items as well. As Gebhardt and Adams showed, the OECD scaling trend results for Germany are biased and overestimate the performance of German students in PISA 2003 and PISA 2006.

12.4 Discussion

This chapter has addressed the question of an adequate trend model for German data from the PISA 2000, PISA 2003 and PISA 2006 studies. In general, trend analyses within a country are a task with different requirements in contrast to comparing countries using cross-sectional results from PISA studies. Different models for scaling trend data and analyzing trends have been discussed with respect to the PISA trend design; in the context of PISA, Gebhardt and Adams (2007) discuss the appropriateness of the original OECD scaling, a concurrent calibration and a conditional analysis of the concurrent calibration for each country in PISA 2000 and PISA 2003. Xu and von Davier (2008) elaborate on different linking models, Monseur and Berezner (2007) examine the effect of omitting link items and Mazzeo and von Davier (2008) discuss the assessment design and analysis models for trends in PISA in comparison to the National Assessment of Educational progress (NAEP) in the USA. They point out, that a very conservative test design with minimum changes over assessment cycles is a key for the stability of trends in NAEP since reporting trends requires more precision in the proficiency distribution estimates than reporting country comparisons.

In this chapter, a national trend model for reporting trends for Germany (Prenzel et al., 2007) and for the German federal states (Prenzel et al., 2008) has been presented and its fit to the German data has been investigated. The model estimates item parameters for German data concurrently for all three cycles and is based on the marginal trends by Gebhardt and Adams (2007). As a result of empirical analyses of factorial validity and item fit analyses, the model was estimated using the link items only for Reading and all available items, link and unique items, for Mathematics and Science. Furthermore, the difference in the underlying construct between the long and the short reading tests in German data became evident and thus the national trend model was restricted to the short test with identical items in all three cycles. This result had not been expected, since the same framework was used to construct both the long and the short tests. The reading assessment was constructed according to a fully detailed framework with a balanced items distribution over subscales. However, the short test does not seem to measure the same construct as the long test in Germany. In contrast, the long versions for mathematics and science do measure the same construct respectively, according to the empirical analyzes, while the assessment instruments were in part constructed according to not yet fully detailed framework versions.

For mathematics and science, the trend results from the national trend model were rather consistent with results from Gebhardt and Adams and with results from the OECD scale values. Only for the domain of reading were the trend results from the OECD report different from the national calibration and Gebhardt and Adams´ marginal trends; these differences may be due to the different test instruments used in the different trend models.

With respect to implications for trend reports from PISA cycles, modeling competencies on the basis of different test instruments over cycles will be a fundamental challenge. Other challenges are obviously variations of item difficulties across countries within a study (country DIF) and across cycles within each country (item drift). An essential prerequisite for providing reliable trends seems to be a test design that administers as many link items as possible in exactly the same set-up within booklets over cycles. This issue is not at the focus of the present chapter and given all restrictions in constructing PISA assessments, this might be the hardest challenge to master. From the perspective of trend analyses, it seems to be of special importance to ensure a construction of test instruments that implement link clusters of items for each domain which are held constant over cycles. Ideally, even the assignment of link clusters to booklets might be kept constant, resulting in link booklets.

Depending on the booklet designs of consecutive PISA cycles, appropriate scaling models for trend analyses have to be developed. One way of thinking might be to accept different models for different questions and to report cross-sectional results on the international PISA scale, while trend results are reported on national scalings. This strategy would provide results with a high degree of fit between data and scaling results; however, implementing it would make it necessary to communicate the rationale for different scaling models to the public.

Another way of thinking might be to relate the cross-sectional scaling and the trend scaling as closely as possible. To address the major difference between national or marginal trend models and original trend or reports from OECD scalings, basing cross-sectional results on the same set of items as trend model results, the items in link clusters only, might be an option. However, a large number of items in the assessment of a major domain would then be omitted in the scaling of proficiency distributions for the combined scales. Instead, these items might then be constructed more independently to assess more distinct subscales or variations from the combined scale. Even if both models, for cross-sectional comparison and trends, were based on the same set of link items only, any country DIF would still be a threat to consistent results for both purposes. However, the rather consistent results for mathematics and science give an indication that country DIF might be a source of much smaller inconsistencies as different test forms (in reading) are. This is one of many questions for future research.

References

Adams, R. J., & Carstensen, C. H. (2002). Scaling outcomes. In R. J. Adams & M. Wu (Eds.), *PISA 2000 technical report*. Paris: OECD.

Adams, R. J., & Wu, M. L. (2002). *PISA 2000 technical report*. Paris: OECD.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370.

Carstensen, C. H, Prenzel, M., & Baumert, J. (2008). *Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt?* [Trend analyses in PISA: How did competencies in Germany develop between PISA 2000 and PISA 2006?] Zeitschrift für Erziehungswissenschaften, Sonderheft 10/2008, pp. 11–34.

Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement, 8*(3), 305–322.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking*. New York: Springer.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results*. Princeton, NJ: ETS.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement, 8*(3), 323–335.

OECD. (2003). *The PISA 2003 assessment framework – Mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.

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

OECD. (2005). *PISA 2003 technical report*. Paris: OECD.

OECD. (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006*. Paris: OECD.

OECD. (2007). *PISA 2006: Science competencies for tomorrow's world* (Analysis, Vol. 1). Paris: OECD.

OECD. (2009). *PISA 2006 technical report*. Paris: OECD.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., et al. (Eds.). (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* [PISA 2006. Results of the third international comparison]. Münster, Germany: Waxmann.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., et al. (Eds.). (2008). *PISA 2006. Die Kompetenzen der Jugendlichen im dritten Ländervergleich*. Münster, Germany: Waxmann.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: University Press.

Wu, M., Adams, R. J., Wilson, M., & Haldane, S. (2007). *Conquest 2.0*. Camberwell, Australia: ACER Press.

Xu, X., & von Davier, M. (2008). Linking for the general diagnostic model. In *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 1, pp. 97–112). Princeton, NJ: IEA-ETS Research Institute.

About the Authors

Domenico Angelone is a researcher at the Institute for Educational Evaluation (Associated Institute of the University of Zurich). Together with Urs Moser, he is responsible for the PISA sampling in Switzerland and for the national and cantonal Swiss PISA reports. Correspondence: Institute for Educational Evaluation, Wilfriedstrasse 32, 8032 Zürich, Switzerland. Email: domenico.angelone@ibe.uzh.ch

Eduardo Backhoff is a professor at the Institute for Educational Research and Development of the University of Baja California, Mexico. His work focuses on large-scale assessment. He developed the first computer-administered high school admission test and directed the national assessment for elementary education in Mexico. He is consultant to projects on the design of PISA and TALIS context questionnaires and the head of a project that examines the feasibility of translating national standard academic tests into the Mayan language. Correspondence: Km. 103 Carretera Tijuana-Ensenada, c.p. 22830; Ensenada, Baja California, México. Email: backhoff@uabc.edu.mx

Werner Blum has been a Professor of Mathematics Education at the University of Kassel since 1975. His current research areas include empirical investigations into teaching and learning mathematics; quality development in mathematics teaching, especially a responsible involvement in the German Education Standards in mathematics for lower and upper secondary levels; national and international comparative studies in mathematics; modelling and applications in mathematics education; proofs and proving. He has been a member of the PISA Mathematics Expert Group since 2000. In 2006, he received the Archimedes Prize of MNU.

Claus H. Carstensen is Professor for Psychology and methods of educational research (University of Bamberg, Germany). He has a training in psychometrics and item response theory and has contributed to scaling and data analysis of PISA 2000 at the international consortium at ACER (Melbourne) as well as of PISA 2000, PISA 2003 and PISA 2006 at the German national project centre. Currently he is involved in the German National Educational Panel Study (NEPS).

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5,

Correspondence: University of Bamberg, 96045 Bamberg, Germany. Email: claus.carstensen@uni-bamberg.de

Luis Angel Contreras-Niño is a professor at the Institute for Educational Research and Development of the University of Baja California, Mexico. He specializes in assessment development. His researcch focuses on assessment paradigms in educational research and the development of elementary and middle school large-scale assesments. He directs a project that evaluates the quality of elementary education in the Mexican state of Baja California. Correspondence: Km. 103 Carretera Tijuana-Ensenada, c.p. 22830; Ensenada, Baja California, México. Email: angel@uabc.edu.mx

John Dossey is Distinguished Professor of Mathematics Emeritus at Illinois State University. He has served as a member of the PISA Mathematics Expert Groups for PISA 2003, 2006, and 2009 and member of the PISA Problem Solving Expert Groups for PISA 2003 and 2012, chairing the 2003 group. He has served as President of the National Council of Teachers of Mathematics (NCTM), Chair of the Conference Board of the Mathematical Sciences (CBMS), Chair of the Mathematical Sciences Advisory Committee of The College Board, and consultant to the National Assessment of Educational Progress, Second and Third International Studies of Mathematics, the SAT and ACT examinations.

Harrie Eijkelhof has a Ph.D. in physics education and is currently director of the Freudenthal Institute for Science and Mathematics Education. He has been in charge of a number of national curriculum development projects, such as PLON, ANW and NLT. His research deals with interdisciplinary science education and the learning of physics. Until recently he was vice-dean of the Faculty of Science in charge of bachelor education.

Andreas Frey is Professor for Research Methods in Educational at the Friedrich-Schiller-University Jena (FSU), Germany and Director of the department of Research in Evaluation and Methodology in the Centrum of Teacher Education and Educational Research. Before joining the FSU, he worked at the Leibniz-Institute for Science and Mathematics Education (IPN) in Kiel, Germany as senior researcher and scientific coordinator of PISA 2006 in Germany. He directs several research projects on computerized adaptive testing, psychometrics and statistical modeling, and is author of numerous publications on CAT and other measurement topics. He is also involved in the development of operational CATs in school settings and co-authored the Multidimensional Adaptive Testing Environment (MATE).

Kylie Hillman is a Senior Research Fellow in the National Surveys research program at the Australian Council for Educational Research, where she currently works on the TIMSS, PIRLS and PISA projects for Australia. She has previously worked on a number of longitudinal studies of young people's progress through secondary school into early adulthood. Correspondence: Australian Council *for* Educational Research, 19 Prospect Hill Road, Camberwell VIC 3124. Email: hillman@acer.edu.au

Eckhard Klieme is a Full Professor of Educational Science at Goethe University and Director of the Center for Research on Educational Quality and Evaluation at the German Institute for International Educational Research (DIPF) in Frankfurt am Main. He has been affiliated with PISA at the national and the international level since the first PISA survey in 2000, e.g. as a consultant to the INES Network A and the PISA Governing Board, as a member of the PISA 2003 Problem Solving Expert Group and the National Project Manager for PISA 2009 in Germany. Recently, as the chair of the PISA 2012 Questionnaire Expert Group and the Study Director for PISA 2015 Questionnaire Development, he made significant contributions to the conceptual foundations and the analytical goals of PISA. His research interests include teaching quality, school effectiveness, and educational measurement. Correspondence: DIPF, Schlossstrasse 29, 60486 Frankfurt am Main, Germany. Email: klieme@dipf.de

Mareike Kobarg is a research fellow at the Leibniz-Institute for Science and Mathematics Education (IPN, Kiel, Germany). Her research focuses on science teaching and learning mainly using video based methods. She has currently published further analysis concerning the international Comparison of Science Teaching and Learning in the OECD countries. Correspondence: Leibniz Institute for Science Education (IPN), Olshausenstr. 62, 24098 Kiel, Germany. Email: kobarg@ipn.uni-kiel.de

Johanna Kordes has a Ph.D. in cognitive psychology and works as a senior educational researcher at Cito Institute for Educational Measurement. She has been involved in the questionnaire development, verification and adaptation processes for PISA 2009. She was a consultant of questionnaire development and national assessment in several consultancy-projects in countries in Europe. Currently, she is part of the SurveyLang-consortium for the European Survey on Language Competencies (ESLC).

Sheila Krawchuk is a Senior Statistician at Westat in the United States, with 23 years of experience in the overall management, planning, design, and development of complex business and social surveys. She has worked on PISA since 1999 and currently manages the sampling and weighting for nearly 70 PISA countries. Before joining Westat, she worked for 9 years at Statistics Canada. Correspondence: Sheila Krawchuk, 3100 Wyntree Drive, Norcross, GA, 30071, USA. Email: SheilaKrawchuk@Westat.com

Ulf Kröhne received his Ph.D. at the Friedrich-Schiller-University Jena, Germany. Since 2009 he works as a senior researcher at the German Institute for International Educational Research (DIPF) in the competence cluster Technology Based Assessment (TBA). He is conducting mode-effect studies for the National Educational Panel Study (NEPS) in Germany. As the first author of the Multidimensional Adaptive Testing Environment (MATE) he is involved in the development of different operational adaptive tests in school settings and in health care.

Christian Monseur is Professor at the University of Liège in Belgium. His main lectures are devoted to quantitative research methods, including sampling theory, measurement models, and statistical analyses. Christian worked for 10 years as a Research Assistant at the University of Liège. From 1999 to the end of 2002, he was a Senior Research Fellow at the Australian Council for Educational Research. He was data manager for the OECD/PISA study in 2000 and project director for the PISA Plus study. He is currently a member of the OECD PISA Technical Advisory Group and that of the OECD PIAAC survey. Correspondence: Départemen Education et Formation, Université de Liège, bld. du Rectorat 5 (B32) 4000 Liège, Belgique. Email: cmonseur@ulg.ac.be

Urs Moser is Director of the Institute for Educational Evaluation (Associated Institute of the University of Zurich) and member of the Swiss National Project Management PISA. Together with Domenico Angelone he is responsible for the additional cantonal samples and assessments in PISA as well as for other large scale assessments in the German Part of Switzerland. Correspondence: Institute for Educational Evaluation, Wilfriedstrasse 32, 8032 Zürich, Switzerland. Email: urs.moser@ibe.uzh.ch

Tarek Mostafa is an economist at the University of London (Institute of Education – Centre for Learning and Life Chances in Knowledge Economies and Societies). He is mainly interested in economic and social policy with a particular emphasis on the economics of education, educational inequalities, and political economy. He used PISA data extensively in comparative analyses of educational inequalities in OECD countries and in the assessment of educational performance and policy. Correspondence: Institute of Education – University of London, (LLAKES Centre). 20 Bedford Way, London WC1H 0AL. England, United Kingdom. Email: T.Mostafa@ioe.ac.uk. Blog: www.tarekmostafa.net

Michael Neubrand is a professor of Mathematics Education at Carl-von-Ossietzky-University in Oldenburg (Germany). He was a member of the German PISA Consortium for PISA-2000 and PISA-2003, when Germany took ample national options, esp. in the domain of mathematics. He was – in PISA-2003 jointly with Werner Blum – the responsible person for the mathematics expert group of PISA in Germany.

Mogens Niss is a Professor of Mathematics and Mathematics Education at Roskilde University, Denmark. His research is focused on mathematical modelling, mathematical competencies, the nature of mathematics education research and assessment of mathematics. He has been a member of the mathematics expert group in PISA since the very beginning. He was the Secretary General of the International Commission on Mathematical Instruction 1991–1998 and is currently a member of the Education Committee of the European Mathematical Society.

Manfred Prenzel is Professor of Empirical Educational Research (Susanne Klatten Endowed Chair) and Dean of the TUM School of Education (Technische Universitaet Muenchen, Germany). He has been a member of the OECD Science Expert Group since the first PISA survey in 2000. As the national project manager for PISA 2003, 2006 and 2012, he supplemented the international surveys in Germany with additional samples and assessments that provided data on the broader range of conditions of teaching and learning on the school and classroom level. Manfred Prenzel is also Director of the German Center for International Large Scale Assessment (ZIB) at TU Muenchen. Correspondence: TUM School of Education, Schellingstr. 33, D 80799 Muenchen, Germany. Email: manfred.prenzel@tum.de

Silke Rönnebeck has a Ph.D. in natural sciences and works as a research fellow at the Leibniz-Institute for Science and Mathematics Education (IPN, Kiel, Germany). She was part of the national coordination team of PISA 2006 in Germany. Her research focuses on the development and assessment of science competencies and the improvement of pre-service and in-service science teacher education based on methods of IBST. Correspondence: Leibniz Institute for Science and Mathematics Education (IPN), Olshausenstr. 62, 24098 Kiel, Germany. Email: roennebeck@ipn.uni-kiel.de

Keith Rust is Vice President and Associate Director of the Statistical Group and Westat, located in Rockville, Maryland, USA. He is also Reseach Professor at the Joint Program in Survey Methodology at the University of Maryland at College Park, where he teaches advanced classes in sample survey statistics. He has been a member of the PISA Consortium and the PISA Technical Advisory Group since the project's inception, and chair of the TAG since 2001. He is a Fellow of the American Statistical Association, an elected member of the International Statistical Institute, and is a former member of the U.S. Committee on National Statistics. Correspondence: Westat, 1600 Research Boulevard RE400, Rockville, MD, 20850-3129, USA. Email: KeithRust@Westat.com

Elwin Savelsbergh is assistant professor and director of the science and mathematics teacher education programmes at the Freudenthal Institute for Science and Mathematic Education. He was on the project team for the national physics curriculum project NiNa. His research focuses on physics problem solving, dynamic systems modeling, and inquiry learning.

Katrin Schöps is a researcher at the Leibniz-Institute for Science and Mathematics Education (IPN, Kiel, Germany). She holds a Ph.D. in Ecology and an M.Sc. in Biology. Her research focuses on the development and assessment of science competencies across the life span with an emphasis on large scale assessments. Correspondence: Leibniz Institute for Science Education (IPN), Olshausenstr. 62, 24098 Kiel, Germany. Email: schoeps@ipn.uni-kiel.de

Nicki-Nils Seitz is statistician (Dipl.-Stat.). He has been data analyst and researcher in the field of medical science and educational research. Since 2007, the focus of his research is the application and the further development of multidimensional adaptive testing (MAT), amongst others using classification methods with MAT. At the moment, he is researcher at the department of Research Methods in Educational at the Friedrich-Schiller-University Jena, Germany.

Guillermo Solano-Flores is Associate Professor at the School of Education of the University of Colorado at Boulder, USA. His work focuses on the development of multidisciplinary approaches that address linguistic and cultural diversity in testing. He is currently involved in projects concerning the semiotic analysis of science and mathemtics items and the cultural adaptation of tests in national and international assessment programs. Correspondence: School of Education, 249 UCB. University of Colorado, Boulder. Boulder, CO 80309-0249. Email: guillermo.solano@colorado.edu

Sue Thomson is the Director of the Educational Monitoring and Research Division and the Research Director of the National Surveys research program at the Australian Council for Educational Research. As the National Research Coordinator for Australia in the Trends in International Mathematics and Science Study (TIMSS), which measures achievement in mathematics and science for students in grades 4 and 8, the Progress in International Reading Literacy Study (PIRLS), which measures reading literacy of grade 4 students, and the National Project Manager for Australia for the OECD Programme for International Student Assessment (PISA), which examines reading, mathematical and scientific literacy of 15-year-old students, Dr Thomson's research at ACER has involved extensive analysis of large-scale national and international data sets as well as analysis of longitudinal data. Correspondence: Australian Council *for* Educational Research, 19 Prospect Hill Road, Camberwell VIC 3124. Email: thomson@acer.edu.au

Ross Turner is a Principal Research Fellow at the Australian Council for Educational Research. He is ACER's project manager for its role in the international consortium that has implemented the 2000, 2003, 2006 and 2009 administrations of the Programme for International Assessment (PISA) survey for the OECD. He manages the PISA mathematics expert group, and leads mathematics test development for PISA.

Daniel Urbach is a Research Fellow at the Australian Council for Educational Research (ACER). At ACER he is primarily responsible for psychometric and data analysis work for various ACER projects and currently reports to the Psychometrics and Methodology Research Program. He has been involved in some of ACER's PISA analysis work which included data cleaning, scaling and trend analyses over a 2 year period. Correspondence: Australian Council for Educational Research (ACER), 19 Prospect Hill Rd, Camberwell, VIC, Australia 3124. Email: urbach@acer.edu.au

Matthias von Davier is a research director in the Research & Development Division at Educational Testing Service in Princeton, NJ, USA. He manages a group of researchers concerned with methodological questions arising in large-scale international comparative studies in education. His current work involves advancing the psychometric methodologies used in analyzing cognitive skills data and background data from large-scale educational surveys, such as the OECD's upcoming PIAAC and the ongoing PISA, as well as IEA's TIMSS and PIRLS. His work also involves the development of software for multidimensional models for item response data, and the improvement of models and estimation methods for the analysis of data from large-scale educational survey assessments. Correspondence: Educational Testing Service, MS 12 T, Princeton, NJ, 08541. Email: mvondavier@ets.org

Kitty Williams is an economist with a particular interest in the economics of education. Over the past 12 years she has held positions with a number of agencies in the Washington area, applying economic principles and econometric methods to several government-sponsored projects in education and health. She maintains an interest in production-function models of educational achievement and, in particular, in questions about the elasticity of substitution with respect to the various resources influencing student achievement.

Trevor Williams is now retired. As an Associate Director at Westat from 1995 through 2009 he managed the day-to-day activities of the U.S. component of several international studies of student achievement under contract to the National Center for Education Statistics. These studies included the International Reading Literacy Study in 1991, PISA in 2000 and 2003, PIRLS in 2001, and TIMSS in 1995 through 2007. He maintains an interest in the analysis of cross-national data, especially the application of structural equation models in the explanation of variation in student achievement within and between countries.

Mark Wilson is a professor of Education at UC, Berkeley. His interests focus on measurement and applied statistics. In the past year he was elected president of the Psychometric society, and also became a member of the US National Academy of Education. He has recently published three books: one, *Constructing measures: An item response modeling approach,* is an introduction to modern measurement; the second, *Explanatory item response models: A generalized linear and nonlinear approach,* introduces an overarching framework for the statistical modeling of measurements; the third, *Towards coherence between classroom assessment and accountability* is about the relationships between large-scale assessment and classroom-level assessment. He has also recently chaired a National Research Council committee on assessment of science achievement—*Systems for state science assessment.* Correspondence: Education, UC Berkeley, Berkeley, CA 94720, USA. Email: MarkW@berkeley.edu