Heidi Harju-Luukkainen
Nele McElvany
Justine Stang *Editors*

Monitoring Student Achievement in the 21st Century

European Policy Perspectives and Assessment Strategies

Springer

Monitoring Student Achievement in the 21st Century

Heidi Harju-Luukkainen
Nele McElvany • Justine Stang
Editors

Monitoring Student Achievement in the 21st Century

European Policy Perspectives and Assessment Strategies

Springer

Editors
Heidi Harju-Luukkainen
Nord University
Levanger, Norway

Justine Stang
Center for Research on Education
and School Development (IFS)
TU Dortmund University
Dortmund, Germany

Nele McElvany
Center for Research on Education
and School Development (IFS)
TU Dortmund University
Dortmund, Germany

ISBN 978-3-030-38968-0 ISBN 978-3-030-38969-7 (eBook)
https://doi.org/10.1007/978-3-030-38969-7

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

About the Contributors

Eva L. Baker is Distinguished Professor of Education at the UCLA Graduate School of Education and Information Studies and Founding Director of the National Center for Research on Evaluation, Standards and Student Testing (CRESST). Her research focuses on the integration of standards, instruction and measurement using technology.

Ildikó Balázsi is Psychometrician and Head of the Department for Analyses of Public Education, Hungarian Educational Authority. She has a master's degree in Applied Mathematics and has been working on international and national assessments for almost 20 years. She was the NPM of PISA 2006, 2009 and 2012 and the NRC of PIRLS 2006 and 2011 in Hungary and has a special interest in policy-relevant analysis of ILSA data.

Sigrid Blömeke is Director and Professor at the Centre for Educational Measurement, University of Oslo (CEMO), Norway. Her research focuses on the assessment of teacher competence and the examination of relations between teacher competence, instructional quality and student achievement, in particular in mathematics.

Michelle Braš Roth has recently retired from the National Centre for External Evaluation of Education, Croatia. Her main activities and responsibilities as Head of the PISA Centre were National Project Manager for the OECD/PISA and TALIS surveys, Croatian Representative to the PISA Governing Board since 2004 and TALIS Board of Participating Countries since 2010, National Research Coordinator for IEA/ICILS 2013, National Coordinator for Education for All 2000 Assessment (UNESCO project) and National Coordinator for Monitoring Learning Achievement (UNESCO/UNICEF project). She started her career as a school counsellor in 1984 and in 1998 moved on to work as senior advisor for general development of the school system at the Croatian Ministry of Science, Education and Sports. Later, she joined the Croatian Education and Teacher Training Agency with the main responsibility in the supervision and in-service training of all primary school principals

and school pedagogues in the field of school leadership, management and development of teaching methods.

Rosalia Castellano is Full Professor of Economic Statistics at the University of Naples "Parthenope" (Italy). She teaches Statistics for Economics and Sample Surveys and Data Quality. Her main publications concern income distribution and inequality, labour market, education, gender gaps, generational mobility and environmental economics.

Alfonso Echazarra is a PISA Analyst at the Organisation for Economic Co-operation and Development (OECD). He obtained his doctorate in Social Statistics from the University of Manchester and has coordinated, written or collaborated on the following reports: PISA 2015 Results: Policies and Practices for Successful Schools, Ten Questions for Mathematics Teachers, Low-Performing Students and several PISA in Focus.

Andreas Frey is Professor of "Educational Psychology: Measurement, Evaluation and Counselling" at Goethe University Frankfurt, Germany, and Professor II at the University of Oslo, Norway. His major research interests are item response theory, computerized adaptive testing and large-scale assessment methods. He was and is involved in several large-scale assessments.

Henrik Galligani Ræder is Doctoral Research Fellow at the Centre for Educational Measurement, University of Oslo (CEMO), Norway. His research focuses on the Norwegian national assessments, vertical scaling and numeracy proficiency of Norwegian students.

Heidi Harju-Luukkainen is Professor of Education at Nord University, Norway. She holds a Ph.D. in Education, Special Education Teacher Qualification and Qualification in Leadership and Management from Finland. She has published more than 150 international books, journal articles and reports as well as worked in more than 25 projects globally. She has developed education programs for universities, has been a PI of PISA sub-assessments in Finland and has functioned as board professional.

Johannes Hartig is Professor of Educational Measurement at the DIPF and Goethe University Frankfurt, Germany. His major research interests are psychometric models for learning outcomes and the effects of learning opportunities. He was and is involved in several large-scale assessments

Dirk Hastedt is the Executive Director of the International Association for the Evaluation of Educational Achievement (IEA). He is also Coeditor in Chief of the IEA-ETS Research Institute's journal *Large-Scale Assessments in Education*. He holds a Diploma in Mathematics from the University of Hamburg and a Dr. Phil. in Education from the University of Vienna.

Miyako Ikeda is Senior Analyst at the OECD, leading PISA data analysis and reporting. She received her master's degree from the University of Tokyo and her doctorate from Columbia University. Before joining OECD in 2004, she was with the IIEP/UNESCO focusing on a regional student assessment (SACMEQ) and also with the World Bank contributing to Vietnam's student assessment.

Stefan Johansson, Ph.D., is Senior Lecturer and Researcher at the University of Gothenburg, Department of Education and Special Education. His research interests centre on questions regarding the validity of International Large-Scale Assessments (ILSAs), teacher assessments and standardized tests. In his previous publications, he has utilized ILSA data to address the use and consequences of ILSA, cross-country achievement differences as well as the role of teacher competence for student achievement.

Dominique Lafontaine is Full Professor of Educational Science at the University of Liège. She has been involved since 1990 in large-scale assessments at both national and international levels and has been a Member of the PISA Reading Expert Group since 1998 and of the Questionnaire Expert Group for PISA 2018. She is an Expert Consultant for national assessments in the French Community of Belgium and a Member of the Compulsory Education Monitoring Board.

Sergio Longobardi is Associate Professor in Economic Statistics at the University of Naples "Parthenope" (Italy). His research interests are especially in the fields of education economics and data quality.

João Marôco is Associate Professor at ISPA-IU and Invited Professor at NOVASBE where he teaches statistics, data analysis and research methods. He is also a Consultant for Educational Assessment and Statistics at the World Bank. Between 2014 and 2018, he was a Member of the Board of Directors of IAVE, I. P., where he coordinated PISA, TIMSS, TIMSS Advanced, PIRLS and ICILS. He is the Author/Coauthor of more than 300 research papers related to statistics, psychometrics and education and 4 books in Statistics and Psychometry. According to Google Scholar, his academic works have been quoted more than 21 000 times (H=51, i10=146).

Nele McElvany is Professor and the Executive Director of the Center for Research on Education and School Development (IFS) at TU Dortmund University and has been a Member of its board of directors since 2010. She heads the working group "Empirical Educational Research with a Focus on Teaching and Learning K-12". Her research is on designing in-depth, cross-sectional and longitudinal studies using questionnaires, tests, video-based observations and experimental designs with large-scale data analyses. Her work on individual, social and institutional conditions of educational processes and outcomes is nationally and internationally presented in conferences and publications.

Harold F. O'Neil Jr. is a Professor of Educational Psychology and Technology at the University of Southern California. His research interests include computer-based teaching and assessment of twenty-first-century skills. His most recent edited book is *Using Games and Simulations for Teaching and Assessment: Key Issues* (2016).

Rolf Vegar Olsen is Professor and Codirector at the Centre for Educational Measurement, University of Oslo (CEMO), Norway. His research focuses on substantive, methodological and policy-related issues of national and international large-scale assessments.

László Ostorics is a Reading Researcher and Head of the Department of Assessment and Evaluation, Hungarian Educational Authority. He has been working in development, implementation and reporting of national and international large-scale assessments such as NABC and PISA since 2006 and has been the National Project Manager of PISA for Hungary since 2013. Currently, he works on the renewal of the Hungarian National Assessment System.

Dragica Pavlović Babić, Ph.D., is Associate Professor in Educational Psychology at the University of Belgrade. She has been the OECD/PISA National Project Manager in Serbia for five cycles. She has gained remarkable experience in leading and supporting the development of the framework for monitoring the inclusiveness in general education as well as the establishment of a national assessment system.

Meyvant Þórólfsson is Assistant Professor at the University of Iceland. He completed his M.Ed. degree from the Iceland University of Education in 2002 with an emphasis on the conceptions of time and space in school mathematics and school science. He obtained his Ph.D. from the University of Iceland in 2013, where he examined the transformation of the science curriculum of Iceland. His main research fields are curriculum theory, assessment of learning, science education, mathematics education and educational research.

Silvia Salchegger is working in the field of international large-scale assessments since 2001. Since 2008, she is a Researcher at the Federal Institute for Educational Research, Innovation and Development of the Austrian School System (BIFIE). Her research interests are school-level tracking, academic self-concept, gender differences and competencies of immigrant students.

Claudia Schreiner has been working in the field of international and national large-scale assessments since 1999. Up to 2008, she was responsible for the national implementation of international studies in Austria (e.g. as PISA NPM). She has been Head of the Department for Education Monitoring and Educational Standards and, from 2014, Director of the Federal Institute for Education Research, Innovation and Development of the Austrian School System (BIFIE). Since 2018, she is working for the University of Innsbruck. Her research interests are competency

orientation and educational standards, (formative) assessment of student performance, equity and evidence-oriented quality development.

Justine Stang is a Postdoctoral Researcher at the Center for Research on Education and School Development (IFS) at TU Dortmund University. She earned her Ph.D. in Psychology in 2017. Her recent work focuses on social and educational psychology research themes like stereotype threat, instructional quality or teacher judgement accuracy. Her research is on cross-sectional and longitudinal studies using questionnaires, tests and experimental designs with large-scale data analyses. Her work is nationally and internationally presented in conferences and publications.

Klaudija Šterman Ivančič is a Researcher at the Educational Research Institute. From 2016 on, she is also a National Project Manager for PISA. Her research interest includes the achievement results in different areas of literacy in large-scale assessment studies and their linkage to personality, motivational, behavioural and socio-emotional predictors in national and international arena.

Urška Štremfel is a Research Associate at the Educational Research Institute. Her research interests include new modes of the EU educational governance and its influence on the national educational space. As part of this, she devotes special attention to the role international large-scale assessments play in shaping Slovenia's national education policy and practice.

Birgit Suchań is working in the field of international large-scale assessments since 1999. She was working as a researcher for PISA until she got National Research Coordinator for PIRLS and TIMSS in 2004. As of 2013, she changed back to PISA where she is currently responsible for the implementation of this project in Austria as the National Project Manager. Her research interests are gender differences, equity in education and teaching structure.

Alejandra Tiana is Professor of Theory and History of Education at the National Distance Education University (UNED) of Spain. Among other responsibilities, he has been the Director of the National Institute for Quality and Evaluation (INCE), Rector of UNED and Secretary of State for Education and Vocational Training. He has authored or coauthored 25 books and more than 200 articles or chapters.

Gunda Tire works at Foundation Innove, government organization under the auspices of the Ministry of Education and Research. She is responsible for implementation of PISA in Estonia. She has been the Editor of three national reports and worked with schools, media and other stakeholders to disseminate PISA in Estonia and beyond.

Liz Twist is Head of Assessment Research at the National Foundation for Educational Research in the UK. She has been National Research Coordinator for PIRLS over three cycles (England, Northern Ireland and Scotland) and is currently

Chief Reading Consultant for PIRLS 2021. She is a Former Primary School Deputy Head Teacher.

Mari-Pauliina Vainikainen is Associate Professor and the Leader of Research Group for Educational Assessment at Tampere University, Finland. She studies the development of thinking skills and transversal competences based on longitudinal large-scale assessment studies. She also coordinates the assessment of the new innovative domains in the Finnish PISA team.

Kajsa Yang Hansen, Ph.D., is Professor of Education at the University of Gothenburg and University West. Her research concerns educational quality and equity from a comparative perspective. Currently, she is investigating if students' opportunities to learn can be associated with the recent educational reforms implemented in Swedish compulsory and upper secondary schools and trying to contribute greater understanding of the long-term trends in Swedish pupils' educational outcomes. She also has an interest in quantitative analytical techniques for large-scale survey data, such as multi-level modelling, structural equation modelling (SEM) and second-generation SEM.

Chapter 1
Introduction to Monitoring Student Achievement in the Twenty-First Century

Heidi Harju-Luukkainen, Justine Stang, and Nele McElvany

Improving quality and effectiveness of investment in education is one of the key objectives of the European Strategic Framework for Education and Training (ET 2020). Educational assessment is a part of this quality assurance system. As countries are moving towards more equitable educational policies and practices, some form of monitoring of the expected progress is required. However, there are challenges with this. Countries and areas have individual ethnic and socioeconomic profiles with multiple and unique mechanisms affecting educational outcome and children's positive learning trajectories, which makes international as well as national comparisons challenging (see closer Chap. 22). Therefore, also local monitoring solutions and regional know-how are in high demand (see, for instance, Garvis et al. 2019). Further, this leads to the need of constant development of not only assessment practices and policies but also research methods in the field of student monitoring. Here, expert competency is desired. Countries are in need of experts that are capable of putting international, national and local educational outcomes into a context and understanding the limitations of them.

How does educational assessment look like in the twenty-first-century Europe? There are naturally different perspectives on how educational assessment is viewed and implemented in different countries across Europe. Also the assessment terminology used in the literature varies and different forms of assessments might have different labels of terms depending on the country's assessment context. These differences are reflected in each country's policy documents, assessment strategies and further also in research conducted on the field. Nevertheless, the assessment field

H. Harju-Luukkainen (✉)
Nord University, Levanger, Norway
e-mail: heidi.k.harju-luukkainen@nord.no

J. Stang · N. McElvany
Center for Research on Education and School Development (IFS), TU Dortmund University, Dortmund, Germany

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_1

has developed rapidly during the last 20 years and the amount of international and national assessments conducted yearly has sky rocketed.

As previously mentioned, local-, national- and international-level assessments are to be monitored. These assessments give policy makers and practitioners the latest information in order to develop and understand their education context. The goal is therefore to get comparative information about educational achievement to improve teaching in learning. However, we know often very little of other countries' assessment policies and practices outside our own. While remedial actions are made and taken with attention on the local context, sometimes an in-depth understanding of, for instance, the long-term consequences or larger global influences is missing. Therefore, a more complex understanding of different educational systems, assessment strategies, policies, practices and their connections is needed. Given that we live in a globalised world, it is important that we understand the context of others in order to reflect our own and also to justify possible actions. With this book we want to challenge policy makers and researchers to explore the complexity and diversity of student monitoring, where various standpoints can be taken in order to improve the quality of the overall system.

This book brings together student assessment academics from across Europe to explore current questions around student monitoring. All of the chapters of this book discuss not only the complexity but also the connections around assessment and policy documents as well as strategies, highlighting shared policies, practices, and also enablers and barriers across them. The chapters are balanced of theory and empirical data to ensure suitability as well as readability for people outside of the discipline of student monitoring. Background information and core empirical results allow for an in-depth view into other nations' assessments.

This book is divided into two parts. The first four chapters focus on general discussions. They provide a general picture of overarching areas of international large-scale assessments that have been of interest in international literature. The book starts with a chapter on **International Student Assessment: Aims, Approaches and Challenges** authored by *Miyako Ikeda and Alfonso Echazarra*. This chapter outlines the aims and approaches of the Programme for International Student Assessment (PISA) undertaken by Organisation for Economic Co-operation and Development (OECD), highlighting similarities and differences compared with assessments conducted by International Association for the Study of Educational Achievement (IEA). It describes the manner in which these international student assessments have helped steer policy dialogue and decisions at the international and regional levels. It concludes with a critical review of current approaches and practices in light of future possibilities, especially in view of future student monitoring. In the third chapter, authored by *Dirk Hastedt*, the focus is turned towards the **History and Current State of International Student Assessment**. The history of large-scale assessments spans more than half a century. However, during the last decades, there has been a substantial change in the public view on student assessments both nationally and internationally. Different countries have developed their monitoring systems rapidly and the importance of different assessments has grown. This chapter critically discusses the current state of the European student monitor-

ing. The fourth chapter **Methodological Challenges of International Student Assessment** is authored by *Andreas Frey* and *Johannes Hartig*. International large-scale assessments are expanding, covering even more population and subject domains than ever before. With the fast development of international as well as national assessments, several challenges have been identified that are connected to the methodology of international student assessments especially within methods and technology available today to collect, scale and analyse data. This chapter describes five current methodological challenges that should be addressed so that large-scale assessments can continue to provide highly useful information on educational outcomes in the future. The fifth chapter authored by *Eva Baker* and *Harold O'Neil* takes us out of Europe to the USA. Both of the authors have been working for several decades with assessment practices across the globe influencing with their work the assessment policies and practices in Europe and the USA. This chapter, **The Assessment Landscape in the United States, From Then to the Future**, shows the parallels of the European and American contexts and shifts the focus from Europe to a global context when it comes to the development of international large-scale assessment practices and policies and the influences within the countries. It finalises our general topics by demonstrating how policies and practices in one relatively large and important economic area can fast change the assessment policies and practices in a global perspective. The goal of this chapter is to substantially review the question of what is the state of assessment in the USA when it comes to accountability purposes in the pre-collegiate (excluding early childhood) educational level as well as briefly discuss the work force contexts. All of the chapters in the first part of the book prepare readers for the second part by giving them a global as well as historic understanding of aims and approaches of student monitoring.

The second part of the book is country specific and in total 15 different countries' assessment systems policies and practices across Europe are presented. The sixth chapter **Monitoring Student Achievement in Austria: Implementation, Results and Political Reactions** is authored by *Claudia Schreiner, Birgit Suchań,* and *Silvia Salchegger*. In this chapter, they describe how measuring and monitoring the outcomes of the Austrian school system has been systematically established only over the past 20 years and, further, how Austria today has an extended system monitoring based in international and national large-scale assessments of multiple age groups. The seventh chapter **Use of Assessments to Inform Educational Policies in French-Speaking Belgium** is authored by *Dominique Lafontaine.* In this chapter, Lafontaine takes a closer look at challenges of French-speaking Belgium when it comes to assessments. Here the national assessments have been developed only lately and there are not yet national assessments developed by professionals that can be used to evaluate trends. Therefore, the only tools available to rigorously evaluate trends for this language area are international assessments. French-speaking Belgium has participated in international assessments since the early seventies, and their results are highly valued by policy makers. The eighth chapter **International and National Assessments in Croatia** is authored by *Michelle Braš Roth.* This chapter discusses the Croatian perspective on assessment. Here, monitoring students' assessment and achievements has become a generator for key changes in

education systems, and it determines significant trends in the development of educational practice in the future. The ninth chapter, **The Evolution of National and International Assessment in England** authored by *Liz Twist,* turns the focus towards England. This chapter introduces some national reading assessments and the country's involvement in international surveys of achievement, focusing on how reading is defined and assessed in PIRLS (the Progress in International Reading Literacy Study) and how this compares to the approach in England's national assessments. This chapter also discusses the future of statutory assessment in England. The tenth chapter, **Educational Assessment in Estonia** authored by *Gunda Tire,* takes a closer look at a system with excellent results and remarkable levels of equity. Interestingly in this chapter, the focus is also turned from national and international assessments to the future of national evaluation system, where more focus is given on formative assessment and use of digital tools which support learning and the development of each learner. The eleventh chapter, **Educational Assessment in Finland** authored by *Mari-Pauliina Vainikainen* and *Heidi Harju-Luukkainen,* is dedicated to Finland. Even though the Finnish education system has received a lot of interest, very little attention has been paid to the model of the Finnish educational assessment system and the lack of standardised measurement and control. Thus, these factors in large are contributing to the overall functioning of the system. In this chapter, the authors provide a historical overview of the development of the assessment model in Finland and further give a description of its current form. **Educational Assessment in Germany** is the twelfth chapter in this book. It is authored by *Nele McElvany* and *Justine Stang*. This chapter presents an overview of Germany's participation in several national and international assessments. It also explains Germany's current educational monitoring system and how it has been affected by results of international student assessments. The thirteenth chapter **The Hungarian Educational Assessment System** is authored by *Ildikó Balázsi* and *László Ostorics*. Since 1968 Hungary has participated in approximately 25 international large-scale student assessments. Participation in these assessments and the development of a national assessment system are intended to inform educational policy makers, professionals and the public. This chapter presents the history and the current state of the Hungarian assessment system with special focus on international studies and the National Assessment of Basic Competencies as its main pillars. The fourteenth chapter, **Educational Assessment in Iceland** by *Meyvant Þórólfsson,* describes the historic development of national and international assessments in Iceland. How the pendulum in a historic perspective has swung from an emphasis on transmission of knowledge to be measured as learning outcomes (products) to an emphasis on learning as a metacognitive activity elevating formative assessment focused on processes of learning rather than products of learning. The fifteenth chapter, **The Long March Towards School and Student Assessment in Italy** authored by *Rosalia Castellano* and *Sergio Longobardi,* describes the challenges as well as benefits of international assessments for the development of a country's education system. In Italy the assessment culture has had considerable difficulty in permeating the Italian school system, and the themes of school evaluation have entered the Italian political agenda

only in the last 15 years, although the Italian participation in international student assessments such as PISA, TIMSS and PIRSL has always been significant. Chapter sixteen, **Large-Scale Assessments in the Norwegian Context**, is authored by *Henrik Galligani Ræder*, *Rolf Vegar Olsen,* and *Sigrid Blömeke***.** This chapter describes the role of international large-scale assessments in Norway as well as presents results from an international perspective. Also the benefits and limitations of the assessment system in its whole, and with its different tools, are discussed against a framework that distinguishes between educational monitoring, support for teaching and learning and certification as core functions of educational assessments. Chapter seventeen, **International Large-Scale Assessments: Trends and Effects on the Portuguese Public Education System,** is authored by *João Marôco***.** This chapter gives a brief overview of the Portuguese scores in major international large-scale assessments and their effects on shaping the Portuguese educational system. It also interestingly frames how OECD suggestions were used to support and justify education policies aimed at the curricula reformulation, teaching practices, students' support programs and schools' management. Chapter eighteen, **International Assessment Studies in Serbia Between Traditional Solutions, Unexpected Achievements and High Expectations** authored by *Dragica Pavlović Babić,* is dedicated to Serbian assessment system. International assessment studies have revealed that educational outcome of Serbia has been low, but the results have not made any visible influence on the curricula, teaching methodology, assessment practice in school and in-service teacher education so far. The nineteenth chapter **Assessment Policy and Practice of Slovenia** is authored by *Klaudija Šterman Ivančič* and *Urška Štremfel.* In this chapter, one of Slovenia's most important goals in the field of education today is discussed. This goal is the establishment of a culture of quality and assessment, which is based on the concept of evidence-based policy, where participation in large-scale assessments plays an important role. In the twentieth chapter, **Monitoring of Student Achievement in Spain** authored by *Alejandra Tiana,* the focus is turned towards Spain. This chapter begins by outlining developments in the monitoring of student achievement in Spain and its current status, before going on to analysing the challenges posed and potential directions for future development. Over the past six decades, the ILSA has changed the landscape of Swedish student assessment in many positive ways; however, it also has identified several areas of problems. In the twenty-first chapter, **Student Assessment in the Landscape of International Large-Scale Studies**, *Kajsa Yang Hansen* and *Stefan Johansson* discuss the assessment policies and practices of Sweden. They start with a retrospective view of the ILSA studies in Sweden and further discuss their benefits and drawbacks for the system.

As a summary, all of the chapters in this book make a valuable contribution to the debate on educational assessment in Europe. Therefore, we would like to invite all practitioners, policy makers and researchers to use this book to understand, design, analyse or evaluate approaches. Only through rich information and shared understanding, the assessment work in Europe can be further developed.

References

European Strategic Framework for Education and Training (ET 2020). Retrieved from https://ec.europa.eu/education/policies/european-policy-cooperation/et2020-framework_en

Garvis, S., Harju-Luukkainen, H., & Yngvesson, T. (2019). Towards a test-driven early childhood education: Alternative practices to testing children. In G. Barton & S. Garvis (Eds.), *Compassion and empathy in educational contexts* (pp. 61–77). Cham: Palgrave Macmillan. https://doi-org.libproxy.helsinki.fi/10.1007/978-3-030-18925-9_4.

Part I

Chapter 2
International Student Assessment: Aims, Approaches and Challenges

Miyako Ikeda and Alfonso Echazarra

Introduction

International student assessments are essential for improving education around the world. They fuel debate and provide powerful information and data to help educators and policy makers identify strengths and weaknesses of their school systems. Assessments serve to raise awareness and calls for accountability in the public eye. Participating countries in the Programme for International Student Assessment (PISA) led by the OECD (Organisation for Economic Co-operation and Development), for instance, have almost tripled over the last 20 years since the first assessment in 2000. Interest in international benchmarking of student performance continues to increase. In 2000, 31 countries participated in PISA, of which 28 were OECD members. Approximately 80 countries and economies took part in 2018 PISA. For PISA 2021 even more have already committed to participate. The other organisation which conducts major international student assessments on mathematics, science and reading is the International Association for the Evaluation of Educational Achievement (IEA). The Trend in International Mathematics and Science Study (TIMSS) was conducted by the IEA in 1995, 1999, 2003, 2007, 2011 and 2015, with another planned for 2019. The Progress in International Reading Literacy Study (PIRLS) by the IEA was conducted in 2001, 2006, 2011 and 2016, with another intended for 2021.

M. Ikeda (✉) · A. Echazarra
Organisation for Economic Co-operation and Development, Paris, France
e-mail: Miyako.IKEDA@oecd.org

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_2

International Student Assessments: An Overview

International student assessments, such as PISA and TIMSS/PIRLS, contribute to education policies and practices at the national level in three important respects. First, they provide reliable and internationally comparable indicators on student performance and other education outcomes and facilitate the monitoring of shifts over time. Second, analysing how international student assessments' data correlate with contextual information can contribute to improve education systems, schools and teacher quality. Finally, international assessments carry great value for the formulation of national policies and practices because they frame current educational debates and highlight internationally agreed-upon metrics and methodologies.

Major international student assessments are generally low stakes for individual students and schools. Survey designs and sampling are typically optimised to obtain results at the country or sub-national level, and assessment results are mainly intended for system-level analysis. Currently, major assessments share similar methodological and implementation steps in achieving their outcomes. First steps include developing a framework and designing appropriate instruments, creating a survey design and sampling plan, and establishing a standardised implementation procedure. These guide the development of operation manuals and various trainings for participants. This is followed by the translation and validation of survey instruments, the drawing of samples and the collection of data. The final steps are the processing and coding of data, the computation of weights, the development of scales for student performance and relevant background data, and the preparation of the database for public access. To maximise data access and information use, detailed publications and technical documents are prepared, published and disseminated online and in print, usually free of charge.

In general, international student assessments aim to collect data to benchmark student performance and to provide comparable indicators across participating countries. Student performance is measured through carefully developed tests based on an agreed-upon assessment framework. Many countries currently create and implement their own national education assessments to measure a variety of domains and interest areas. However, these assessments rarely provide results that allow for a direct and comprehensive international comparison. By participating in international student assessments, countries can compare their students and education systems directly with others. Participation in these assessments over a number of cycles also allows for the monitoring of trends and country performance over time. While current international assessments measure similar outcomes, they also look at different aspects of these outcomes. For instance, PISA focuses on the level of student preparedness for full participation in a society while TIMSS and PIRLS focus more on the level of student mastery of the school curriculum.

International student assessments, in addition to providing data on student performance, also collect contextual information on students, schools and school systems. Because data gathered in such assessments are limited in their capacity to

identify causal inferences, policy makers and educators must determine how best to improve student performance in their own countries. By correlating the contextual information with student performance, however, it is possible to identify student or school groups at risk. Correlational information also help in the examination of education policies and practices shared by high-performing students, schools and countries.

While results of international student assessments help policy makers, researchers and educators identify strengths and weaknesses of a given education system, such assessments provide valuable data and information for countries to learn from one another. For example, the topic that most interests researchers who use PISA data at this time is that of equity (Hopfenbeck et al. 2018). Internationally comparable indicators on equity have shown to be of great interest to them, more specifically the relationship between student socio-economic status and their academic performance. By showing that this relationship exists in essentially all participating countries, policy makers and researchers can use these results to better understand why this relationship is weaker or stronger in certain education systems. The debate on the possible trade-off between equity and overall education quality (e.g. average student performance) continues to be hotly disputed. However, PISA has shown that high education results and equity can be achieved by identifying those countries that have achieved both at the same time (OECD 2010, 2013a, 2016). Background contextual information collected also helps countries understand the roles of school organisation, teaching strategies and practices, the learning environment or parental support. In correlating background information with a variety of education outcome indicators, educators and policy makers can identify target groups that need further support and those policies and practices which are related to the outcomes. This is invaluable for policy makers and educators who must plan, adjust, implement and pursue effective and impactful policies and practices.

A number of education issues become more salient when education systems are held in comparison. The practice of grade repetition is an illuminating example because education systems around the world handle grade repetition and the challenges of diverse student populations differently. Some systems encourage or require students to repeat a grade if they are deemed unprepared for advancement. Other school systems allow students to advance automatically into the next grade regardless of performance and/or behaviour. In comparing rates of grade repetition across countries, educators can better understand how the quality and equity of their education systems are related to grade repetition. In this light, PISA results have shown that 15-year-old students are spread across wider range of grade levels in those education systems featuring grade repetition. Data shows that overall performance in these systems tend to be lower while the impact of socio-economic status on student learning outcomes is higher in those systems that feature more frequent grade repetition (OECD 2010). These PISA findings have consequently contributed to shape national policies on grade repetition. Belgium, France, Portugal or Spain, for example, introduced new policies that reduced grade repetition, lowering rates over recent years (OECD 2018a).

In addition to the important data results gained from international assessments, the frameworks themselves, the methodologies applied and the instruments designed also carry great value since they provide shared points of reference to policy makers, educators and researchers. Available to everyone online and/or in print, assessment frameworks provide the current definition of constructs and metrics to measure student performance. Detailed technical documents outline and describe survey design and methodologies, sampling, instrument development scaling approach, translation and verification processes, survey operation, and database structure and management. Those who are involved in developing and administering national assessments for their own countries often refer to these materials for insight and comparison, as well as for ideas and direction in determining their approach and their methodology. This has helped improve national assessments and fostered synergies between international and national assessments. This also sheds light on the distinct nature of each national assessment.

International student assessments have helped steer policy dialogues internationally and regionally. At this time, internationally comparable indicators on education contribute to the monitoring of the Sustainable Development Goals adopted by the United Nations in September 2015. More specifically, its fourth goal seeks to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all. At the regional level, the European Union set objectives for its member education systems that applied PISA indicators as benchmarks (OECD 2013b; European Commission 2018). One target is to reduce the share of underachieving 15-year-old students in reading, mathematics and science to less than 15% by the year 2020.

Approaches to International Student Assessment: PISA Case Study

Every 3 years since 2000, PISA has assessed the extent to which 15-year-old students near the end of compulsory education have acquired the knowledge and skills that are essential for full participation in modern societies. PISA aims to gauge whether students can reproduce the knowledge they have acquired in and outside of school. It also determines whether students can extrapolate from what they have learned and apply knowledge and skills in unfamiliar contexts. To maintain relevancy and impact, PISA has consistently broadened its assessment of competencies and dispositions with each assessment cycle.

Whereas the mastery of subjects such as mathematics, science and reading and their application are regarded by all as being essential, other skills and dispositions are also recognised for their importance. In response, PISA has included a new domain and/or topics with each of assessments. PISA 2003 included student self-assessment of their learning strategies. For 2006, PISA incorporated an assessment of student attitudes towards science. Both PISA 2003 and 2012 featured assessment sections concerning problem-solving skills. PISA 2012 also offered countries the possibility

of measuring financial literacy. In 2015, PISA assessed student ability to solve problems collaboratively. For PISA 2018, the assessment features a section concerning 15-year-old student capacity to examine local, global and intercultural issues, to understand and appreciate the perspectives and world views of others and to engage in open, appropriate and effective interactions with people from different cultures.

While PISA's main objective remains to provide reliable and comparable measurements of student performance internationally, PISA also collects contextual data that describe school systems and other important aspects. Such data help policy makers and educators improve and raise their national performance standards because they show a granular picture, helping establish relationships between student performance and a range of factors and influences, such as family background, student attitudes towards learning, their habits and their life in and outside of school. The assessment also surveys principals about the staff and material resources in their schools, aspects of school management and funding, the school's curricular emphasis, any extracurricular activities offered and the general context of instruction. Questionnaires for parents and teachers are also available for countries that are interested in learning more from their perspectives.[1]

How PISA Differs from Other International Student Assessment Studies

Current international student assessments differ among one another in a number of key areas. For example, the most distinctive between PISA and TIMSS/PIRLS concerns what exactly is being measured. Wagemaker (2008) identified that the assessments embodied the differences in their histories and the aims of the two organisations. From its inception, PISA was created to monitor the extent to which students near the end of compulsory schooling had acquired the knowledge and skills essential for full participation in society. This aim falls within the broader mission of the OECD and embodies its mandate. When PISA was launched in 1997, the OECD initiated the Program Definition and Selection of Competencies (DeSeCo) to develop a conceptual framework which would define key competencies to guide the assessment. Through DeSeCo, the OECD collaborated with a wide range of scholars and specialists of a broad range of disciplines, as well as input from country representatives. They established that PISA results would contribute to valued outcomes for societies and individuals, help individuals meet important demands in a wide variety of contexts and be significant not only for specialists but for all individuals (OECD 2005). This was accomplished by identifying specific challenges and values common across countries and cultures, as well as acknowledging the diversity in values and priorities. Shaped by the DeSeCo framework for key competencies, the

[1] In PISA teacher data are linked to student data at the school level, but not at the individual student level.

first PISA assessment in 2000 took place, focusing predominantly on student competencies in the domains of reading, mathematics and science. The framework established a foundation and pathway for how additional competency domains, which are essential for student success in life, could be incorporated into future assessments (Rychen and Salganik 2003).

The IEA creates assessments focus instead on understanding the linkages between the intended curriculum (what policy requires), the implemented curriculum (what is taught in schools) and the achieved curriculum (what students learn), drawing on the concept of 'opportunity to learn' (https://www.iea.nl/our-studies). TIMSS and PIRLS are formulated and designed to focus on the teaching/learning process and to assess the extent of knowledge acquired by students after a fixed period of schooling. IEA's interests lie 'in addressing questions of efficiency and equity with respect to the ability of educational systems to deliver what is mandated by the curriculum' (Wagemaker 2008).

PISA differs from those of the IEA in their sampling approach and questionnaire content. PISA applies an age-based sampling with a target population of 15-year-old students who are in grade 7 or above. In contrast, studies by the IEA apply grade-based sampling, regardless of student age. For example, the TIMSS grade 8 target populations are defined as 'all students enrolled in the grade that represents 8 years of schooling counting from the first year of ISCED Level 1, providing the mean age at the time of testing is at least 13.5 years' (LaRoche et al. 2016). In terms of the content coverage of questionnaires, IEA survey questions emphasise school curriculum. For example, in those countries which taught science as separate subjects at the grade 8 level (e.g. biology, chemistry, physics and earth science), the TIMSS 2015 survey had students respond to questions specific to each subject, in addition to other aspects of the curriculum and their home and school lives in general. Also, IEA studies had a dedicated questionnaire on curriculum that was completed by national research coordinators from participating countries that is specifically designed to collect information on the national contexts for learning (Hooper 2016). In contrast, the PISA 2015 student questionnaire focused on core subjects (e.g. science) and collects comparatively more information on non-academic outcomes (e.g. career expectations, well-being) and other contextual information concerning student life and well-being. Furthermore, country representatives complete a system-level questionnaire that feature education policy questions on such things as teacher training and support, the structure of the education system and school administrative aspects.

Challenges for the Future of International Student Assessments

Organisations leading international student assessments must ensure relevancy for policy makers, researchers and educators who aim to improve education. To fulfil expectations and address changing needs, international large-scale assessments, and

PISA in particular, are facing numerous challenges. An obvious challenge for international student assessments is managing time constraints. The 3-year PISA cycle provides timely information for countries but also requires the organisations, institutions and experts involved in the collection process to closely coordinate for a timely delivery of results. In addition, any innovation in the design and delivery of international student assessments must be balanced with the response burden for students and other education stakeholders. As with any international project, student assessments also face significant financial challenges. Organisations must keep operation costs, which are typically borne by taxpayers in participating countries, at reasonable levels. Beyond these obvious challenges, there are at least six other challenges for organisations leading international-level student assessments.

Responding to Economic, Social and Technological Changes

The competencies students required to succeed in world today are evolving at an ever-increasing rate. The Internet has changed the way people connect with one another and how individuals access information. People and markets are now interconnected globally in ways unimaginable only decades earlier. The labour market is increasingly seeking non-routine, interpersonal and higher-order skills (OECD 2013c; Frey and Osborne 2017). In response, the PISA assessment has also changed over time with the addition of new content areas and the use of technologies that have made the assessment process more streamlined and flexible. New assessment domains have included problem-solving, digital literacy, collaboration and global competence. Because the framework and questions of traditional domains (reading, mathematics and science) are updated every 9 years, the challenge is introducing new content areas while maintaining a consistency in traditional domains to measure trends over time.

The PISA assessment is now computer based rather than paper-pencil in a majority of countries. This change in 2015 came in response to the rise of digital literacy and the manner in which digital technology has been incorporated into daily life. A computer-based assessment is expected to expand the potential range of how questions can be presented to students and how responses could be given.

Despite these innovation efforts, more is needed to ensure the long-term relevance of international student assessments. Vital skills, such as creativity, entrepreneurship or communication skills, could be considered for assessment, as well as a number of traditional school subjects, such as art, history, geography and music. The ability of students to communicate in a foreign language is another important area that could be measured internationally. In this regard, the European Commission assessed the foreign language skills in 16 education systems in 2011, and a foreign language assessment is being considered as part of PISA 2024. Assessing these new domains for more than half a million students will certainly be a challenge fraught with complexity and debate. However, the inclusion of such subject domains helps

ensure the relevancy and worthiness of international student assessments in our changing world.

Making Assessment Relevant for All

An impressive number of education systems are currently participating in international student assessments. For instance, nearly 60 countries participated in TIMSS 2015 and nearly 80 countries and economies are taking part in PISA 2018. An increased number of education systems greatly enhances the value of benchmarking. However, the participation of middle- and low-income countries brings new challenges. The most significant is ensuring that international student assessments can accurately measure the knowledge, skills and learning contexts of a complex diversity of student populations. To address this challenge, the OECD initiated *PISA for Development* in 2014, a project designed to incorporate middle- and low-income countries into the main PISA assessment. This project adds more items at the lower end of the performance distribution, creates survey instruments that are more relevant to the context of these countries and offers countries the possibility of sampling out-of-school children. It also helps countries in survey implementation and national report development. Together, these initiatives have proven extremely beneficial and are being progressively incorporated into the main 2018 and 2021 PISA assessments. Greater flexibility and continual adjustments are still required to make international large-scale assessments equally relevant to all countries, particularly to middle-income countries.

Making Results Useful for All Stakeholders

Critics of international student assessments have highlighted that such projects appear to mainly serve policy makers and researchers. They argue that this has limited the relevancy of international student assessments for other important groups of education stakeholders, namely, schools, parents and teachers (Carabaña 2015). An OECD report evaluating the policy impact of PISA (OECD 2008) in fact cited that a majority of policy makers and researchers reported to be knowledgeable about PISA whereas only a third of parents and school principals reported similarly. To some extent, the relevance of international student assessments is limited by their design because they are not intended to present data at the individual classroom, school or even school district level or to have the results providing direct feedback to participating students and schools. To remedy this, linking international student assessments to national or regional assessments, either by having a subset of students taking both assessments or including some items from international student assessments in national assessments, can position students or individual schools on international scales and within the framework of international standards. This would

immediately raise the awareness of international student assessments, their role in international benchmarking, and increase the potential of raising learning standards.

In the case of PISA, whose natural target audience are policy makers, great efforts continue to be made in reaching out to other stakeholders who might benefit from its results by adjusting questionnaire content.

For example, the *PISA for Schools project*—in which a group of schools is invited to participate voluntarily—provides direct feedback to individual schools on the abilities and learning opportunities of their students using PISA as a benchmark. PISA has created special publications for other stakeholders, such as for parents (*Let's Read Them a Story*), for teachers (*Ten Questions for Mathematics Teachers* and *How PISA Can Answer Them; Qudwa: Global Teachers' Forum*) and for those who are interested in specific areas such as environment, gender and digital technology in education (*Green at Fifteen?*; *The ABC of Gender Equality; Students, Computers and Learning*).

Improving Test and Questionnaire Reliability

The test design and scaling procedures in international student assessments undergo regular update to improve and to incorporate current advances in the field. PISA 2015, for instance, increased the number of common items, transitioned from a paper-based to a computer-based assessment and applied a more flexible statistical model for scaling (e.g. two-parameter model or 2PL). The persisting challenge to improve reliability, however, lies in the cross-cultural comparability of questionnaire scales, most notably that concerning scalar invariance (i.e. whether the average of a certain indicator can be compared across cultures). PISA is addressing this issue in a number of ways, including working closely with field and technical experts, triangulating data sources whenever possible and being innovative in the questionnaire design. These include anchoring vignettes, using forced-choice items, reversed keyed items, including a reference point, and the use of various item formats in the field trial. In late 2018, in response to the clear need, the OECD gathered specialists and researchers from around the world for a conference that focused on novel approaches in the fields of measurement equivalence and invariance testing.

Drawing Causal Inferences

There is a recognised place of international student assessments in evaluating education systems and in the formulation of evidence-based policy decisions. However, the drawing of causal inferences from cross-sectional observational studies is problematic (Rutkowski and Delandshere 2016). PISA reports carry a disclaimer stating clearly that PISA cannot identify cause-and-effect relationships between policies/practices and student outcomes. Researchers can reduce the uncertainty

around causal inferences if analytical methods that rest on specific underlying assumptions are applied. These include using propensity score analysis (Hogrebe and Strietholt 2016; Kaplan 2016), instrumental variables (Pokropek 2016), applying a difference in differences approach (Rosén and Gustafsson 2016) or conducting cross-subject analysis with student-fixed effects (Bietenbeck 2014; Echazarra et al. 2016; Schwerdt and Wuppermann 2011).

In general, design of international student assessments can better accommodate causal analyses by integrating experimental and longitudinal studies within their frameworks. For instance, they can facilitate post-testing to measure the effectiveness of educational interventions or encouraging more countries to follow sampled students into the future. Countries that have implemented longitudinal studies following PISA studies include Australia, Canada, Denmark and Switzerland (OECD 2018b). Questionnaires focusing on collecting more information on all assessment subjects would increase the number of possible cross-subject analyses with student-fixed effects. Questionnaires could also include items that are retrospective with the aim of collecting data on the cumulative experience of students. This would provide a more holistic picture rather instead of a simple snapshot of student experience. Video studies within the international student assessment structure have proven to aid in better understanding classroom practice (Cuban 2013) with one example being the 1999 TIMSS Video Study. The upcoming TALIS (*Teaching and Learning International Survey*) Video Study will also look into the classroom dynamics in eight countries.

Enhancing Transparency and Communication

Organisations leading international student assessments have always prioritised transparency and communication. In every cycle, PISA makes the database, frameworks and questionnaires publicly available, explaining all technical aspects in a dedicated report (OECD 2017). It also presents the results in multiple formats (e.g. reports, country notes, PISA in Focus, blogs, working papers, slides, infographics). For illustrative purposes, PISA also releases a number of actual test questions. PISA is a collaborative effort in which the OECD and countries are supported by many actors, including international and national experts and specialised contractors. The governance structure of PISA requires a broad range of actors to participate, contribute and help in the design and implementation of the project. Rigorous technical standards guide activities at all stages of the assessment. Yet a surprising number of education stakeholders, most notably parents, teachers and principals, remain unaware of the process behind international student assessments, often viewing these important projects with suspicion. For instance, in an evaluation of the impact of PISA (OECD 2008), only a small share of parents, principals and media and business representatives reported to be aware of the manner in which the PISA assessment was planned, coordinated and implemented in participating countries. Clearly, organisations like the OECD and the IEA can improve in the way they inform the

public about their international student assessments. More effectively informing the public and their stakeholders about their projects, their purposes and their benefits will demystify the assessments and improve transparency. Simplifying pathways to the information on assessment designs and supporting explanatory and technical materials will broaden their reach and ensure that the data can contribute in meaningful and impactful ways.

Conclusion

International student assessments have done much to help improve education around the world. The OECD and IEA are but two major organisations that currently formulate, design and implement these assessments. Shaped by the missions of their organisations, PISA and TIMMS/PIRLS collect data for different purposes. To maintain relevancy, international student assessments must evolve with societal and public needs while remaining rigorous and robust. Researchers and policy makers have recognised the contribution that international student assessments have made to improve our knowledge base of education. However, organisations leading these assessments must work more deliberately to give all stakeholders the ability to fully access their results. If the right people employ the results of international student assessments correctly, student learning around the world can improve immensely.

References

Bietenbeck, J. (2014). Teaching practices and cognitive skills. *Labour Economics, 30*, 143–153.

Carabaña, J. (2015). *La inutilidad de PISA para las escuelas*. Madrid: Los Libros de la Catarata.

Cuban, L. (2013). *Inside the black box of classroom practice: Change without reform in American education*. Cambridge, MA: Harvard Education Press.

Echazarra, A., Salinas, D., Méndez, I., Denis, V., & Rech, G. (2016). How teachers teach and students learn: Successful strategies for school. In *Education working papers 130*. Paris: OECD Publishing.

European Commission. (2018). *Education and training monitor 2018*. Luxembourg: Publications Office.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280.

Hogrebe, N., & Strietholt, R. (2016). Does non-participation in preschool affect children's reading achievement? International evidence from propensity score analyses. *Large-scale Assessments in Education, 4*(1), 2.

Hooper, M. (2016). Developing the TIMSS 2015 context questionnaires. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015*. Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, MA.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J. A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research, 62*(3), 333–353.

Kaplan, D. (2016). Causal inference with large-scale assessments in education from a Bayesian perspective: A review and synthesis. *Large-Scale Assessments in Education, 4*(1), 7.

LaRoche, S., Joncas, M., & Foy, P. (2016). Sample design in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015*. Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, MA.

Organisation for Economic Co-operation and Development. (2005). *The definition and selection of key competencies: Executive summary*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2008). *The policies that inspired PISA – and the policies that PISA inspired: 2009 report on the evaluation of the policy impact of PISA, OECD document EDU/PISA/GB(2008)35/REV2*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2010). *PISA 2009 results: What makes a school successful? Resources, policies and practices* (Vol. IV). Paris: PISA/OECD Publishing.

Organisation for Economic Co-operation and Development. (2013a). *PISA 2012 results: Excellence through equity: Giving every student the chance to succeed* (Vol. II). Paris: PISA/OECD Publishing.

Organisation for Economic Co-operation and Development. (2013b). *OECD skills outlook 2013: First results from the survey of adult skills*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2013c). *OECD reviews of evaluation and assessment in education: Synergies for better learning*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: PISA/OECD Publishing.

Organisation for Economic Co-operation and Development. (2017). *Technical report*. Paris: PISA/OECD Publishing.

Organisation for Economic Co-operation and Development. (2018a). *Education policy outlook 2018: Putting student learning at the centre*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2018b). *Equity in education: Breaking down barriers to social mobility*. Paris: PISA/OECD Publishing.

Pokropek, A. (2016). Introduction to instrumental variables and their application to large-scale assessment data. *Large-scale Assessments in Education, 4*(1), 4.

Rosén, M., & Gustafsson, J. E. (2016). Is computer availability at home causally related to reading achievement in grade 4? A longitudinal difference in differences approach to IEA data from 1991 to 2006. *Large-scale Assessments in Education, 4*(1), 5.

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-scale Assessments in Education, 4*(1), 6.

Rychen, D. S., & Salganik, L. H. (Eds.). (2003). *Key competencies for a successful life and well-functioning society*. Gottingen: Hogrefe Publishing.

Schwerdt, G., & Wuppermann, A. C. (2011). Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review, 30*(2), 365–379.

Wagemaker, H. (2008). Choices and trade-offs: Reply to McGaw. *Assessment in Education: Principles, Policy & Practice, 15*(*3*), 267–278.

Chapter 3
History and Current State of International Student Assessment

Dirk Hastedt

The Beginning

Empirical pedagogical research has been conducted for a long time and while it is debatable where the start should be identified, it is most likely in the eighteenth century or even earlier. However, large-scale educational assessments started to become prominent in the 1960s. In June 1959, researchers from various disciplines came together and met at the UNESCO Institute for Lifelong Learning in Hamburg. They were interested in exploring the feasibility of a study of educational achievement in different countries and cultures. At that time, it was not clear if this would be achievable at all – something that is taken for granted today. So, one of the two main purposes of the first study was "To discover the possibilities and difficulties attending a large-scale international study" (page 8 in Foshay et al. 1962). Consequently, from the perspective of the researchers involved, the most important outcome was not content related but establishing whether or not it was possible to conduct a quantitative international study on educational outcomes.

Also in this report, Foshay coined the often-cited statement, "If custom and law define what is educationally allowable within a nation, the educational systems beyond one's national boundaries suggest what is educationally possible. The field of comparative education exists to examine these possibilities" (page 7 in Foshay et al. 1962). This shows also their perspective on these studies. International large-scale assessments are about comparing education systems, their outcomes and their relationship with various input factors in order to understand and improve education. They are not intended or conceived to be about ranking and competition between education systems despite the impression that one might sometimes get from todays' discussions and reporting in newspapers and other mass media.

D. Hastedt (✉)
International Association for the Evaluation of Educational Achievement (IEA), Amsterdam, The Netherlands
e-mail: d.hastedt@iea.nl

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_3

This first study marked the foundation of the International Association for the Evaluation of Educational Achievement (IEA) and consequently of international large-scale assessments in education in general. The IEA was legally founded as a non-profit scientific association in Belgium in September 1967 (https://www.iea.nl/legal-status).

After the successful pilot study, several other studies in various subjects and for various age groups were launched. The first pilot study was on mathematics achievement. This subject was chosen because there was a growing interest at that time in improving mathematics and science education, and analyses of current curricula were already completed. Another major reason was that the researchers perceived mathematics to be the subject with the lowest cultural and language influence and therefore the easiest to assess across countries, as Postlethwaite wrote in the study report, "…since the symbols of arithmetic and mathematics are, with trifling exceptions, international problems of semantic and language would be reduced" (page 24 in Postlethwaite 1967).

Developments from the 1980s Onwards

The 1980s saw an increase in IEA studies. Not only were the IEA's Second International Mathematics Study (SIMS) and the Second International Science Study (SISS) conducted but, with the Pre-Primary Project (PPP), the age cohorts assessed were expanded. At this time new areas of investigation were developed including the Written Composition and the Computers in Education (COMPED) studies.

In the beginning of international large-scale assessments, researchers from all participating countries and various disciplines discussed all aspects of the study, sometimes with significant disagreements. Studies were conducted more as coordinated national assessments, but this changed in the 1980s. From then on, all studies were conducted under the leadership of an international study center that was responsible for the international coordination of the study but also for all technical aspects like sampling, scaling, and coordination of the development of the assessment instruments. The way in which assessments in the participating countries were conducted was increasingly standardized across countries, but still not across studies.

During this period, the study centers for the IEA studies were in different research institutes in different countries. Neville Postlethwaite at the University of Hamburg led the SIMS, SISS, and Reading Literacy Study. The Written Composition study was coordinated at the Institute for Educational Research, University of Jyväskylä, Finland; the COMPED study at the University of Twente, Enschede, the Netherlands; the Pre-Primary Project at HighScope Educational Research Foundation; the Civic Education (CIVEd) at Humboldt University of Berlin; and the Second Information Technology in Education Study (SITES) jointly by the University of Twente,

Enschede, the Netherlands, and the University of Hong Kong. Exchange between studies happened at the annual IEA General Assemblies.

The IEA Reading Literacy Study can be seen as a milestone for international large-scale assessments. Based on the experience of previous studies SIMS and SISS, there was a concern about the lack of comparable data from some countries. In response to this uncertainty, a strategy is established to develop a standardized data entry and capture system, including the development of international code books which not only prescribed the format for data entry but also introduced more rigorous quality control procedures. Countries could make national adaptations, but they had to document these changes and submit data in the specified format. This was a major improvement and consequently marks the start of international well-documented databases. Figure 3.1 gives an overview of the IEA studies conducted so far.

In 1994, building on the technical knowledge and tools developed during the IEA Reading Literacy Study, IEA established a Data Processing Center in Hamburg to be responsible for the technical aspects of IEA's international large-scale assessments. Tasks included data processing, the development of software for data entry, and study management and were enlarged later to scaling and sampling and weighting as well as manual preparation and reporting table production. Also a permanent secretariat of IEA was created establishing the IEA as a more professional organization independent of individual studies.

Another landmark for international studies was reached with the Trends in International Mathematics and Science Study (TIMSS). The first cycle was conducted in 1995 and preparations began in 1993. The study center for TIMSS was established at Boston College and first lead by Al Beaton and Ina Mullis who have been heavily involved in the National Assessment of Educational Progress (NAEP). NAEP was first launched in the USA in 1964, and many of the techniques applied in today's international large-scale assessments were developed by researchers working for NAEP. This includes matrix sampling, jackknife procedures, and scaling procedures including usage of plausible values (Beaton et al. 2011). Many of these techniques were introduced in TIMSS and later to all IEA and non-IEA assessments including the Organization for Economic Development (OECD's) Programme for International Student Assessment (PISA).

TIMSS is also a landmark in international large-scale assessments since it marks a change from one-time assessments to studies that are designed as trend studies and which are conducted on a regular cycle. In its first appearance in 1995, TIMSS was the third international mathematics and science study which measured achievement of grades three and four and grades seven and eight students. After SIMS and SISS, the IEA decided to create a study combining these assessments and, since both mathematics and science were the subjects of two previous studies, TIMSS was seen as the third study in these areas. After the success in finding very interesting, and sometimes surprising outcomes, there was an interest in measuring the achievement of the grade 4 students 4 years later with instruments linked to the grade seven and eight assessment of TIMSS 1995.

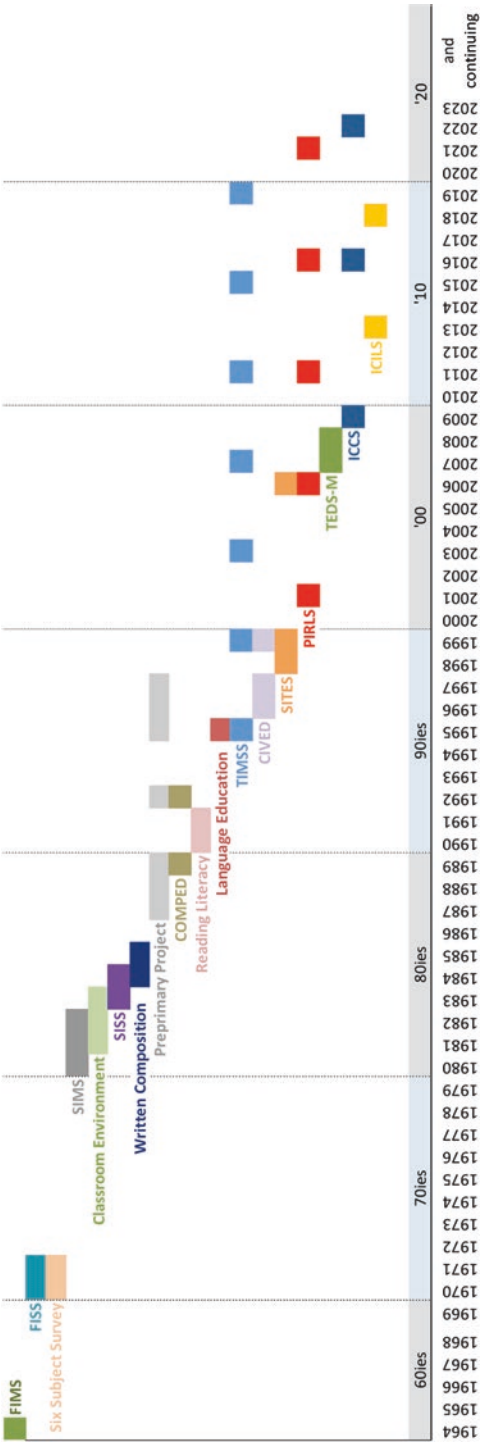

Fig. 3.1 IEA studies conducted up to 2018 and planed in the coming years
Legend: *FIMS* First International Mathematics Study, *FISS* First International Science Study, *SIMS* Second International Mathematics Study, *SISS* Second International Science Study, *COMPED* Computers in Education Study, *TIMSS* Trends in International Mathematics and Science Study, *CIVED* Civic Education Study, *SITES* Second Information Technology in Education Study, *PIRLS* Progress in International Reading Literacy Study, *ICCS* International Civic and Citizenship Education Study, *ICILS* International Computer and Information Literacy Study; bars indicate the years of the main surveys

The TIMSS-Repeat study was conducted in 1999 and cemented the desire to measure students' achievement repeatedly because of the interesting trend results it generated. Questions raised included: how does achievement in various countries develop over time and can any consequences of changes in educational policies be measured? This change has impacted most international assessments, both within and outside of IEA. The development of cyclic studies that are planned to measure trends over time and consequently are now seen by many internationally as fundamental to monitoring the performance education systems.

The increased awareness of differences in learning outcomes after a fixed period of schooling stimulated debates among educators in the OECD countries which then began thinking about launching their own student assessment programs with the aim to monitor student achievement over time. The approach of the OECD is somewhat different and explained below.

This coincided with what may be termed the empirical turn in education (see below). However, measuring achievement multiple times and linking the achievement and background results from one cycle to a previous cycle posed several new challenges and requirements for international large-scale assessments. How can materials be released publicly to demonstrate the content of the assessment but without compromising our ability to measure trends across cycles with linkable assessment instruments? Do items retain their statistical properties across cycles? How can linkage errors between cycles be estimated?

Rankings

The advent and use of international ranking is one of the most contentious outcomes of the pursuit of the comparisons of educational achievement and its antecedents. While all current international large-scale assessments publish league tables in one form or the other (ICCS studies, ICILS 2013, PISA studies, TALIS 2013, PIAAC, TIMSS and PIRLS studies), this was not an emphasis in the first studies. Researchers were interested in the effects of student, teacher and school background as well as system level decisions on students' achievement (see, e.g., Foshay et al. 1962 or Postlethwaite 1967), but not whose students are performing best or worse. Interestingly, Foshay et al. (1962) reported in table 1 of the Twelve Country Pilot Study the differences in variance of student achievement across the different achievement domains but not the mean achievement.

For SIMS, the IEA did not publish league tables in international reports (Garden 1987; Jaji 1986; Livingstone 1986; Pelgrum et al. 1986; Westbury and Travers 1990). However, some newspapers did create and publish league tables from the study data. The problem was that the way in which the rankings were calculated was incorrect. That led to the decision that the IEA Reading Literacy study international rankings were to be published in the international report (table 3.1 in Elley 1992): not because the researchers found them necessarily very enlightening, but to avoid

having wrong international rankings published by others. Since then, all IEA studies and all other international large-scale studies publish league tables in their reports.

However, it should be noted that there are clear differences in the reports of different organizations. The IEA publishes international rankings for average student achievement together with the range, but the emphasis is on understanding the students' performance in a much wider context resulting in most tables in the international reports being ordered by the percentage of students with certain background characteristics (e.g., students speaking the language of test at home or students attending safe schools) or alphabetically (e.g., on trend results). In contrast, most tables in the OECD PISA reports are ordered by mean student achievement or another variable of interest like the achievement gain over 3 years. The first PISA 2000 report (OECD 2001) includes 29 figures including study results by country. In eight of the figures the countries are ordered by achievement result; in seventeen, the countries are ordered by another outcome variable of interest; and in two, the countries are ordered alphabetically. In contrast the just previously published TIMSS 1999 report on mathematics achievement (Mullis et al. 2000) from IEA includes 93 exhibits showing individual country results. In only 3 exhibits the countries are ordered by achievement, in 43 by another variable of interest, and in 47 the countries are ordered alphabetically. The report PISA 2015 in focus (OECD 2018) includes 12 tables and figures with individual country results; in seven of them the countries are ordered by achievement outcomes and in five by other outcome variables of interest.

The Empirical Turn

While the first international large-scale assessments were mostly of academic interest, this somewhat changed in the 1990s. The first studies were mostly seen as fundamental research to understand educational processes across countries and cultures. Or, as Foshay et al. (1962) phrased it, "If custom and law define what is educationally allowable within a nation, the educational systems beyond one's national boundaries suggest what is educationally possible." Policy makers and the general public were not too interested in these studies. Consequently, researchers working in the field spent significant amounts of time writing proposals to donor organizations and ran these studies under severe financial pressure.

In the 1990s the picture changed significantly. Several countries included empirical studies in education in their policies. For example, in Germany, the Constance Resolution of the standing committees of the education ministers of the states in their 1997 meeting was seen as the change in paradigm (Overesch 2007). International organizations were also interested in educational outcomes. Improving educational outcomes was not only seen as a measure to leverage future perspectives of individuals but of economic systems or countries. Changes in educational outcomes could be used in predicting financial developments of countries: an

important factor, for example, for banks giving (or not giving) loans to different countries.

Developments from 2000 Onwards

In the year 2000, the first cycle of OECD's PISA was conducted with the results released in 2001. Previously, the OECD made use of results from other organizations in their main educational publication "education at a glance." For example, the 1997 release of education at a glance (OECD 1997) published input variables in the education sector like percentage of students catered for or the number of teachers from member countries from external sources like the United Nations. The chapter of educational outcomes also drew on results from the IEA Reading Literacy Study (OECD 1997). Then the OECD was requested by its members to conduct their own study of educational outcomes in part as a response to the Nation at risk report in the USA (National Commission on Excellence in Education 1983).

PISA has been conducted every 3 years since 2000. PISA assesses the learning domains of reading, mathematics, and science for 15-year-old students. Until PISA 2012, one of the domains was assessed as a major domain and the other two as minor domains (and therefore covered with fewer assessment materials and consequently measured less precisely). From PISA 2015 on the design is changed, now giving all domains equal emphasis in each cycle.

PISA has helped change the landscape significantly. International student assessments receive much more public interest, probably due the huge investments in public outreach. They also attract the attention of policymakers since the OECD is a governmental organization. For the first time, international large-scale assessments in education moved from educational researchers to an organization dedicated to the economic development of countries.

This location of PISA in an organization with a strong economic focus may be one of the reasons for researchers from the field of economics to be increasingly using educational data in their analyses. Models developed in the field of economics are now also been applied to international achievement studies (see, e.g., Hanushek and Wößmann 2006). Some of the models from the field of economics try to derive causal relationships. Causal modelling uses statistical models that treat data from ILSAs as quasi-experimental data which are then used to draw causal inferences. These modelling techniques include instrumental variable analysis, regression discontinuity analysis, or propensity score matching. These models are based on relatively strong assumptions that cannot be proven, but their plausibility can be made evident. For example, the propensity score matching assumes that all relevant variables are captured in the model. The instrumental variable approach, on the other hand, assumes that the instrumental variable has no indirect effect on the outcome. For details of causal modeling, the reader is referred to Robinson (2014) or Rutkowski (2014).

PISA and IEA Studies: Different Paradigms

An overview of current studies cannot exclude a comparison of the different approaches taken by the OECD in PISA and the IEA studies. PISA and IEA have clearly different approaches rooted in the different aims of the organizations. "The IEA is an international cooperative of national research institutions, government research agencies, scholars and analysts working to evaluate, understand and improve education worldwide"(IEA website https://www.iea.nl/about-us). So, the aim of IEA studies is targeted to understanding education systems and to evaluating the input and output and processes of the education taking place. Consequently, IEA studies follow a curriculum model. Data is collected and analyzed about the intended curriculum (what students should learn), the implemented curriculum (what happens in classrooms), and the achieved or attained curriculum (what students have learned). The IEA makes labor-intensive efforts to contrast students achievements in an area to what students should have learned. All test items are discussed with all participating countries during the test construction. Differential Item Functioning (DIF) analysis is conducted to find and eliminate items from the scaling that have deviating properties in different countries. Finally, a test curriculum matching analysis is conducted and countries scored and compared on each set of items that each country deems to be covered in their curriculum (see, e.g., table B1 in Beaton et al. 1996).

In contrast, as a governmental organization, the OECD has the mission "…to promote policies that will improve the economic and social well-being of people around the world" (OECD website http://www.oecd.org/about/). Consequently, OECD's PISA focus is "How well are young adults prepared to meet the challenges of the future? Are they able to analyse, reason and communicate their ideas effectively? Do they have the capacity to continue learning throughout life? Parents, students, the public and those who run education systems need to know the answers to these questions" (foreword of the first PISA report, OECD (2000)). "OECD/PISA differs from some other assessment programmes in that it is not primarily an assessment of the extent to which students have mastered bodies of knowledge and skills identified in school curricula. It is not an assessment of achievement in school reading, mathematics and science only" (page 14 in OECD 2000). So, in contrast to current IEA studies, PISA does not refer to the participating countries' curricula and the education process taking place in schools but rather focuses on the knowledge and abilities at a certain age level learned **in school or outside** of schools.

These different approaches also impact the definitions of the target populations for IEA studies and PISA. Since the IEA is focused on the curriculum of countries and teaching and education processes, IEA studies are grade based. TIMSS assesses grade 4, grade 8, and grade 12 students, PIRLS grade 4 students only, and ICILS and ICCS grade 8 students only. This is done since curricula are specified in all countries for different grade levels but not based on students' age. In contrast, PISA targets 15-year-old students which makes analysis of school level factors more difficult since countries differ in terms of the age when students enter schools but also

their policies with respect to grade repetition and the opportunities to learn various topics.

The PISA approach also means that students from different grade levels are sampled. A sample of students from each school who are 15 years of age are sampled – irrespective of their grade or classes. IEA studies usually sample intact classes to make the test administration less disruptive in schools and enables the assignment of questionnaires to the subject teachers of the sampled class, be it mathematics, science, or reading. Therefore, in the analysis of TIMSS or PIRLS data, student achievement outcomes can be directly linked to the subject teacher, which is not possible in PISA. A further excurse on the differences can also be found in Wagemaker (2014).

There are also differences between IEA and the OECD with respect to study organization, and study structures as well as the reporting and technical implementations (e.g., different item response models used in scaling the achievement and background data).

Current Participation

Today, nearly all European countries participate in one or more of the international large-scale studies from IEA or OECD as can be seen from Table 3.1. It should also be noted that other organizations conduct studies in education. In Europe, the European Commission has also launched studies in the field of education. These studies collect mostly system level information. An exception is the European Study on Foreign Languages (ESLC) conducted in 2011 in 16 European countries (see http://ec.europa.eu/dgs/education_culture/repository/languages/policy/strategic-framework/documents/language-survey-final-report_en.pdf). There is an increased interest, and possibly also a pressure, from the public as well as intergovernmental organizations on countries to participate in these studies with the effect that participation is rising. Rizvi and Lingard discuss the mechanisms and dynamics in their research quite extensively (Rizvi and Lingard 2006, 2010).

Future Outlook

In 2015, the United Nations declared the sustainable development goals (SDGs) which set targets and goals to be reached by 2030 (https://www.un.org/sustainabledevelopment/sustainable-development-goals/). The SDG 4 targets goals related to education. Compared to the UN's previous millennium goals which were to be reached by 2015, the SDGs are more output oriented. While the millennium goals set the target that primary education should become universal, the SDGs specify that, by 2030, all girls and boys should "complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes"

Table 3.1 Membership and participation of European countries in IEA studies and PISA

Country	IEA member	OECD member	study participation					
			TIMSS grade 4	TIMSS grade 8	PIRLS	ICCS	ICILS	PISA
Austria	Yes	Yes	1995, 2007, 2011, 2019	1995	2005, 2011, 2016	2009		Since 2000
Belgium (Flemish)	Yes	Yes	2003, 2011, 2015, 2019	1995, 1999, 2003, 2011	2006, 2016	2009, 2016		Since 2000
Belgium (French)	Yes	Yes		1995	2006, 2011, 2016			Since 2000
Bosnia and Herzegovina	No	No	2019	2007				
Bulgaria	Yes	No	2015, 2019	1995, 1999, 2003, 2007	2001, 2006,2011, 2016	2009, 2016		Since 2000 (not 2003)
Croatia	Yes	No	2011, 2015, 2019	2019	2011	2016	2013	Since 2009
Cyprus	Yes	No	1995, 2003, 2015, 2019	1995, 1999, 2003, 2007, 2019	2001	2009		Since 2012
Czech Republic	Yes	Yes	1995, 2007, 2011, 2015, 2019	1995, 1999, 2007	2001, 2011, 2016	2009	2013	Since 2000
Denmark	Yes	Yes	2007, 2011, 2015, 2019	1995	2006, 2011, 2016	2009, 2016	2013, 2018	Since 2000
England	Yes	Yes	1995, 2003, 2007, 2011, 2015, 2019	1995, 1999, 2003, 2007, 2011, 2015, 2019	2001, 2006, 2011, 2016	2009		2015
Estonia	Yes	Yes		2003		2009, 2013		Since 2006
Finland	Yes	Yes	2011, 2015, 2019	1999, 2011, 2019	2011, 2016	2009, 2016	2018	Since 2000
France	Yes	Yes	2015, 2019	1995, 2019	2001, 2006, 2011, 2016		2018	Since 2000
Germany	Yes	Yes	2007, 2011, 2015, 2019	1995	2001, 2006, 2011, 2016	2016[c]	2013, 2018	Since 2000
Greece	Yes	Yes	1995	1995	2001	2009		Since 2000

(continued)

Table 3.1 (continued)

Country	IEA member	OECD member	study participation					
			TIMSS grade 4	TIMSS grade 8	PIRLS	ICCS	ICILS	PISA
Hungary	Yes	Yes	1995, 2003,2007, 2011, 2015, 2019	1995, 1999, 2003, 2007, 2011, 2015, 2019	2001, 2006, 2011, 2016			Since 2000
Iceland	Yes	Yes	1995	1995	2001, 2006			Since 2000
Ireland	Yes	Yes	1995, 2011, 2015, 2019	1995, 2015, 2019	2011, 2016	2009		Since 2000
Italy	Yes	Yes	1995, 2003, 2007, 2011, 2015, 2019	1995, 1999, 2003, 2007, 2011, 2015, 2019'	2001, 2006, 2011, 2016	2009, 2016	2018	Since 2000
Latvia	Yes	Yes	1995, 2003, 2007, 2019	1995, 1999, 2003	2001, 2006, 2016	2009, 2016		Since 2000 (not 2012)
Liechtenstein	No	No				2009		
Lithuania	Yes	Yes	2003, 2007, 2011, 2015, 2019	1995, 1999, 2003, 2007, 2011, 2015, 2019	2001, 2006, 2011, 2016	2009, 2016	2013	Since 2006
Luxembourg	Yes	Yes			2006	2009	2018	Since 2000 (not 2012)
Macedonia	Yes	No	2019	1999, 2003, 2011	2001, 2006			2000, 2015
Malta	No	No	2011, 2019	2007, 2015	2011, 2016	2009, 2016		2009, 2015
Moldova	No	No	2003	1999, 2003	2001, 2006			2009, 2015

(continued)

Table 3.1 (continued)

Country	IEA member	OECD member	study participation					
			TIMSS grade 4	TIMSS grade 8	PIRLS	ICCS	ICILS	PISA
Netherlands	Yes	Yes	1995, 2003, 2007, 2011, 2015, 2019	1995, 1999, 2003	2001, 2006, 2011, 2016	2009, 2016	2013	Since 2000
Northern Ireland	No	Yes	2011, 2015, 2019		2011, 2016			2015
Norway	Yes	Yes	1995, 2003, 2007, 2011, 2015, 2019	1995, 2003, 2007, 2011, 2015, 2019	2001, 2006, 2011, 2016	2009, 2016	2013	Since 2000
Poland	Yes	Yes	2011, 2015, 2019		2006, 2011, 2016	2009	2013	Since 2000
Portugal	Yes	Yes	1995, 2011, 2015, 2019	1995	2011, 2016		2018	Since 2000
Romania	Yes	No	2011, 2019	1995, 1999, 2003, 2007, 2011	2001, 2006, 2011			Since 2006
Russian Federation	Yes	No	2003, 2007, 2011, 2015, 2019	1995, 1999, 2003, 2007, 2011, 2015, 2019	2001, 2006, 2011, 2016	2009, 2016	2013	Since 2000
Scotland	Yes	Yes	1995, 2003, 2007	1995, 2003, 2007	2001, 2006			2015
Serbia	No	No	2011, 2015, 2019	2003, 2007				
Slovak Republic	Yes	Yes	2007, 2011, 2015, 2019	1995, 1999, 2003	2001, 2006, 2011, 2016	2009	2013	Since 2003
Slovenia	Yes	Yes	1995, 2003, 2007, 2011, 2015	1995, 1999, 2003, 2007, 2011, 2015	2001, 2006, 2011, 2016	2009, 2016	2013	Since 2006

(continued)

Table 3.1 (continued)

Country	IEA member	OECD member	study participation					
			TIMSS grade 4	TIMSS grade 8	PIRLS	ICCS	ICILS	PISA
Spain	Yes	Yes	2011, 2015, 2019	1995, (2003, 2007)[a], (2019)[b]	2006, 2011, 2016	2009		Since 2000
Sweden	Yes	Yes	2007, 2011, 2015, 2019	1995, 2003, 2007, 2011, 2015, 2019	2001, 2006, 2011, 2016	2009, 2016		Since 2000
Switzerland	No	Yes		1995, 1999		2009	2013	Since 2000
Turkey	Yes	Yes	2011, 2015, 2019	1999, 2007, 2011, 2015, 2019	2001		2013	Since 2000
Ukraine	No	No	2007	2007, 2011				

[a]Basque only
[b]Madrid only
[c]North Rine-Westphalia only

(http://sdg4monitoring.uis.unesco.org/). Compared to the millennium goals, the SDGs require students not only to be in school but also to learn and reach minimum proficiency benchmarks. This represents a major shift in focus from quantity – numbers like enrolment – to quality of education.

This revised focus addresses one of the key problems of some countries' response to the millennium goals where they increased school enrollment without the necessary corresponding investment in supporting infrastructure. This meant that although there were more children in school, what students learned and the number of students reaching minimum proficiency standards decreased. For example, Uganda saw an "…increase of primary school enrolment figures from 2.7 million pupils in 1996 to 5.3 million in 1997, and to 7.1 million in 2005" (Nakabugo et al. 2006), leading to class sizes of more than 70 students in many schools. To avoid these unwanted consequences, the SDGs require countries to monitor educational outcomes such as, for example, the percentage of students reaching a minimum benchmark in literacy and numeracy at the end of primary education. Since the reported data is supposed to be internationally comparable, it is expected that international large-scale assessments like IEA studies or PISA will become increasingly relevant, especially for low- and middle-income countries where so far participation has been lower.

There is also an increasing demand for moving from correlation to causation. Policy makers want to know that if they change an input A in an education system, it will result in an outcome B. This push for causal relationships is not new. Even the report of IEA's first international mathematics study included: "Although the first object of any inquiry of this kind must be to find evidence of association there is a further, more difficult, question. When evidence of association has been found how is it to be interpreted? Evidence of association is necessary if causal relations are to be inferred, but it is not enough" (page 131 in Postlethwaite 1967).

The statistical modelling approaches to achieve causality are explained above. The main problem with these approaches is that they are based on relatively strong assumptions as reported by Rutkowski (2016). But another direction that has gained an increasing interest is to design studies in a way that they are able to model causality. In this respect, the opportunities available with longitudinal studies are increasingly popular since it is assumed that they are getting closer to causality.

While more robust empirical studies are being developed, there are also those studies which may provide a richer and more nuance understanding of the context in which instruction takes place. Together with an interest in longitudinal research, video studies are increasingly popular. The first TIMSS in 1995 included a video study as a study component with mathematics lessons in Germany, Japan, and the USA recorded and analyzed (Stigler et al. 1999). The OECD also launched a video study connected to their Teaching and Learning International Study (TALIS) to be conducted in 2018 in nine countries (http://www.oecd.org/education/school/TALIS-2018-video-study-brochure-ENG.pdf). Video studies are not only quite costly, especially with respect to analysis, but they also create challenges with respect to data protection and personal rights when the video material shall be made publicly available. Instead of making videos from the studies available, exemplary videos demonstrating the main findings were recorded and publicized for example in the TIMSS video study.

IEA and OECD studies currently follow a pseudo-longitudinal approach. This means that, instead of following individual students, student populations are measured repeatedly every 3, 4, or 5 years. Since TIMSS measures grade 4, grade 8, and grade 12 students every 4 years, the same cohort of students is followed, but not individual students. Some countries also include national studies and followed individual students. For example, in Germany TIMSS 2007 students were followed in a longitudinal approach as well as a complementary sample of German grade 9 students who were sampled in PISA 2012. Some national educational large-scale assessment projects also employ a longitudinal design. Probably the largest in Europe is the German National Educational Panel Study (NEPS) that follows different cohorts of students. "As valuable as these cross-sectional studies are, they can only be viewed as isolated snapshots documenting a specific point in the life course at a fixed moment in time. To transform such snapshots into a moving picture, the NEPS will be following individuals over time so that we will be able to reconstruct how competencies unfold over the life course, how competencies and decision-making processes relate to various critical transitions in educational careers, along with how and to what extent they are influenced by the family of origin and the

structure of teaching and learning processes in Kindergarten, school, vocational training, higher education, and later (working) life" (https://www.neps-data.de/en-us/projectoverview/aimsoftheproject/longitudinaldata.aspx).

Although longitudinal studies are usually more expensive due to the tracking and assessment of individual students and despite the fact that longitudinal study data is not comparable to experimental study designs, there is the belief that the data facilitates analysis that gets researchers closer to causal relationships. Consequently, researchers and policy makers increasingly ask for longitudinal studies. There is a great chance that this will also feed into the design of international large-scale assessments in the future.

Other tendencies in today's discussions around ILSAs reveal interests in broadening the scope of international studies by including social science or arts and music to get a more comprehensive picture of education. Also a focus on students' skills rather than their knowledge can be seen. Tertiary education as well as vocational education are also areas that researchers and policy makers mark as interesting extensions. The IEA has approached the tertiary education with its study on mathematics teachers' education, TEDS-M. On the other hand, there appears to be a growing concern among policy makers, students, teachers, and parents in many countries, particularly in Europe, that too many assessments are taking place and leaving too little time for actual teaching.

Lastly, computer-based assessments (CBA) can be expected to play an increasing role in student assessments. CBAs have the advantage that data is available immediately after the assessment, assessments can be more efficient – especially when adaptive systems are implemented – and the collection of process data like the response times enables additional analysis. Furthermore, areas like application of knowledge or cooperation can be easily assessed, and since computers are increasingly used in teaching and learning, CBAs are becoming more natural with respect to students' daily experiences.

It will be interesting to see what the future will bring and the direction that international large-scale assessments will take.

References

Beaton, A. E., Mullis, I. V., Martin, M. O., Gonzales, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years. IEA's third international mathematics and science study*. Chestnut Hill: Boston College.

Beaton, A. E., Rogers, A. M., Gonzalez, E., Hanly, M. B., Kolstad, A., Rust, K. F., Sikali, E., Stokes, L., & Jia, Y. (2011). *The NAEP primer (NCES 2011–463)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Elley, W. B. (1992). *How in the world do students read? IEA study of Reading literacy*. The Hague: IEA.

European Commission. (n.d.). *First European survey on language competences*. Retrieved from http://ec.europa.eu/dgs/education_culture/repository/languages/policy/strategic-framework/documents/language-survey-final-report_en.pdf

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962). *Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg: UNESCO Institute for Education.

Garden, R. A. (1987). *Second IEA mathematics study*. Sampling report. Washington: Center for Education Statistics.

Hanushek, E. A., & Wößmann, L. (2006). Does early tracking affect educational inequality and performance? Differences-in differences evidence across countries. *Economic Journal, 116*(510), C63–C76.

IEA. (n.d.). *About us*. Retrieved from https://www.iea.nl/about-us

Jaji, G. (1986). *Second international mathematics study. The use of calculators and computers in mathematics classes in twenty countries: A source document*. Washington: Center for Education Statistics.

Lehmann, R. H. (2010). Die nationale und internationale Bedeutung empirischer Bildungsforschung. In J.-D. Gauger & J. Kraus (Eds.), *Eine Veröffentlichung der Konrad-Adenauer-Stiftung e. V. Empirische Bildungsforschung: Notwendigkeit und Risiko* (pp. 21–39). Sankt Augustin: Konrad-Adenauer-Stiftung.

Leibniz Institute for Educational Trajectories. (n.d.). *NEPS – National Educational Panel Study*. Retrieved from https://www.neps-data.de/en-us/projectoverview/aimsoftheproject/longitudinaldata.aspx

Livingstone, I. D. (1986). *Second international mathematics study. Perceptions of the intended and implemented mathematics curriculum*. Washington: Center for Education Statistics.

Mullis, et al. (2000). *TIMSS 1999 International Mathematics Report. Findings from IEA's repeat of the Third International Mathematics and Science Study at the eighth grade*. Boston College, US.

Nakabugo, M. G. et al. (2006). Instructional strategies for large classes: Baseline literature and empirical study of primary school teachers in Uganda. *International Journal of Educational Development, 26*(1) https://home.hiroshima-u.ac.jp/cice/wp-content/uploads/publications/report2/AA/Kampala_Uganda.pdf

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. *The Elementary School Journal, 84*(2), 112–130. https://doi.org/10.1086/461348.

OECD. (1997). *Education at a glance: OECD indicators 1997*. Paris: OECD Publications.

OECD. (2000). *Measuring student knowledge and skills*. OECD.

OECD. (2001). *Knowledge and skills for life*. OECD.

OECD. (2018). *PISA 2015 results in focus*. Paris: OECD.

OECD. (n.d.-a). *About the OECD*. Retrieved from http://www.oecd.org/about/

OECD. (n.d.-b). *TALIS 2018. Video study and global video library on teaching practices*. Paris: OECD. Retrieved from http://www.oecd.org/education/school/TALIS-2018-video-study-brochure-ENG.pdf

Overesch, A. (2007). *Wie die Schulpolitik ihre Probleme (nicht) löst: Deutschland und Finnland im Vergleich. Internationale Hochschulschriften: Bd. 492*. Münster: Waxmann.

Pelgrum, W. J., Eggen, T., & Plomp, T. (1986). *Second international mathematics study. The implemented and attained mathematics curriculum: A comparison of eighteen countries*. Washington: Center for Education Statistics.

Postlethwaite, N. T. (1967). *School organization and student achievement. A study based on achievement in mathematics in twelve countries*. Stockholm: Almqvist & Wiksell.

Rizvi, F., & Lingard, B. (2006). Globalization and the changing nature of the OECD's educational work. In H. Lauder, P. Brown, J.-A. Dillabough, & A. H. Halsey (Eds.), *Education, globalization, and social change* (pp. 247–260). Oxford: Oxford University Press.

Rizvi, F., & Lingard, B. (2010). *Globalizing education policy*. New York: Routledge.

Robinson, J. P. (2014). Causal inference and comparative analysis with large-scale assessment data. In L. Rutkowski, M. V. Davier, & D. Rutkowski (Eds.), *Statistics in the social and*

behavioral sciences series. Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 521–545). Boca Raton: CRC Press.

Rutkowski, L. (2016). Introduction to special issue on quasi-causal methods. *Large-Scale Assessments in Education, 4*(1), 1.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study. Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States*. Washington: NCES.

UN. (n.d.). *About the sustainable development goals*. Retrieved from https://www.un.org/sustainabledevelopment/sustainable-development-goals/

UNESCO. (n.d.). *Sustainable Development Goal (SDG) 4*. Retrieved from http://sdg4monitoring.uis.unesco.org/

Wagemaker, H. (2014). International large-scale assessments: From research to policy. In L. Rutkowski, M. v. Davier, & D. Rutkowski (Eds.), *Statistics in the social and behavioral sciences series. Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 11–36). Boca Raton: CRC Press.

Westbury, I., & Travers, K. (1990). *Second international mathematics study. Studies*. Urbana: University of Illinois.

Chapter 4
Methodological Challenges of International Student Assessment

Andreas Frey and Johannes Hartig

Introduction

International large-scale assessments (ILSAs) are the pivotal type of study when it comes to comparing student competencies at an international level. ILSAs are considered to be "remarkably successful in comparing student performances and curricula" (van de Vijver et al. in press, p. 17) and "here to stay" (Singer et al. 2018, p. 77). One key factor in the success of ILSAs lies in their rigorous methodological and psychometric basis. All major ILSAs use current state-of-the-art methodology and, additionally, stimulate methodological innovations, which can then be directly applied at an international level. We currently do not see any serious problems limiting the feasibility of ILSAs. Nevertheless, because education systems worldwide are subject to rapid changes, ILSAs obviously need to evolve as well in order to maintain their capability of providing highly useful information. This applies in particular to the current period, which is characterised by extensive changes due to the introduction of digital technology in education. Against this background, it is necessary to set the course now so that future methodological challenges that can already be anticipated now can be met.

In this chapter, we describe five current methodological challenges of ILSAs and we outline pathways for how to tackle them. By discussing the challenges *adopting new constructs*, *consideration of performance heterogeneity, measurement and statistical modelling of context variables*, *transparency of data and methods*, and *validation of test score interpretations,* we address issues concerned with the collection, measurement, analysis, and usage of ILSA data.

A. Frey (✉)
Goethe University Frankfurt, Frankfurt, Germany

Centre for Educational Measurement, University of Oslo, Oslo, Norway
e-mail: frey@psych.uni-frankfurt.de

J. Hartig
Leibniz Institute for Research and Information in Education (DIPF), Frankfurt, Germany

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_4

Our remarks are based on the assumption that the core objectives of ILSAs will continue to be benchmarking and monitoring in the future and that the descriptive character of ILSAs will thus be retained. This seems to be the most likely development track for ILSAs, although the desire to derive causal conclusions from ILSAs has already been discussed (e.g. Kaplan and Kuger 2016; Singer et al. 2018).

Adopting New Constructs

In today's fast-moving world, which is also characterised by the penetration of digital technology into all areas of life, the social relevance of educational constructs is also in flux. On the one hand, constructs change (e.g. reading competency due to the introduction of digital reading devices). On the other hand, new constructs come into focus and quickly acquire social relevance. For example, the competent use of computers and the Internet is without doubt an important prerequisite for successful participation in today's modern societies. Correspondingly, skills in using information and communication technology (ICT; International ICT Literacy Panel 2002) have been under discussion as an educationally relevant construct for around 20 years now. The relevance of the construct for ILSAs is expressed by the fact that it has already been focused on in the International Computer and Information Literacy Study (ICILS; Fraillon et al. 2013). It can be assumed that new constructs will continue to come into focus in the future. In this respect, there is already a trend towards constructs in which tasks have to be solved cooperatively by several persons (e.g. Herborn et al. 2018). The measurement of such constructs is associated with considerable challenges regarding test development, scoring, and scaling. The main challenge when incorporating such constructs into ILSAs will be to use digital technologies in such a way that the constructs are operationalised as realistically as possible while at the same time satisfying psychometric requirements.

Fortunately, ILSAs are well equipped to identify and adopt new constructs because they have a differentiated international expert panel structure (van de Vijver et al. in press). ILSAs are likely to be faster and more flexible than education systems in this regard. However, most ILSAs still need to implement guidelines on when and how constructs will not be continued.

Consideration of Performance Heterogeneity

The heterogeneity of student performance within and between countries is a persisting challenge for ILSAs that has not yet been fully addressed. At the onset of ILSAs, only paper-based administration was possible. At that time, the practice of using test booklets with item difficulties roughly aligned with the assumed student performance distribution was the best option. However, due to the increased possibilities of using computers for testing purposes, performance heterogeneity can now be handled

much better. Most importantly in this regard, computers can be used to adapt the difficulty of the presented items to the assumed or the measured individual performance level. This brings statistical and psychological improvements compared to assigning test booklets randomly.

The statistical advantages arise from the fact that the statistical information provided by the response to one single item is at its maximum when the difficulty parameter of this item is equal (in the Rasch model) or close to (in other item response theory [IRT] models) to the performance level of the test taker. Adapting the item difficulties of presented items to the performance level can therefore lead to substantial increases in measurement precision.

Regarding the psychological advantages, empirical evidence based on data from the Programme for International Student Assessment (PISA) suggests that the *ability-difficulty fit* (the difference between the ability level of a student and the average difficulty of the items this student worked on) affects test-taking effort and boredom/daydreaming significantly (Asseburg and Frey 2013). Individuals whose ability level was much lower than the average difficulty of the items they had to answer reported lower levels of test-taking effort and higher levels of boredom/daydreaming compared to students whose ability level was equal to or higher than the average item difficulty. The adaptation of the difficulty level of the presented items to the individual student performance, and thus keeping the ability-difficulty fit constant across students, circumvents the potentially performance-reducing effects of the test instrument itself that are mediated by variables such as test-taking effort or boredom/daydreaming. As a result, all tested students have the same opportunity to demonstrate their abilities.

The first steps towards adjusting the difficulty of the presented items to student abilities have already been taken. Starting with the 2009 assessment, PISA offered low-performing countries the option of including item clusters with easy items in the assessment (Organisation for Economic Co-operation and Development 2012a). This mild adjustment of the average item difficulty, however, showed only marginal advantages regarding bias and the root mean square error (RMSE) of ability estimation in the simulation study of Rutkowski et al. (2018). The Progress in International Reading Literacy Study (PIRLS) uses a similar approach to that of PISA. Here, a less difficult version of the regular assessment, called PIRLS Literacy, is available for low-performing countries (Mullis and Martin 2015).

Moving one step ahead, the adaptation of item difficulty to performance level is currently not only carried out on the level of countries but also on the level of individual students in PIAAC (Programme for the International Assessment of Adult Competencies) and PISA. In both ILSAs, multistage testing (MST) was used (for PIAAC, see Organisation for Economic Co-operation and Development 2013a; for PISA, see Yamamoto et al. 2018b). PIAAC used both, computer-based assessment and paper-based assessment. For computer-based assessment, MST was applied. The MST design included several layers of adaptation. Adaptation took previous student responses to cognitive items and background information (education level, native vs. non-native speaker) into account while controlling for item exposure rates and other constraints. Using this MST design proved to be 10–30% more efficient

for the literacy scale and 4–31% more efficient for the numeracy scale compared to a nonadaptive linear test design (Yamamoto et al. 2018a). In PISA, the efficiency gains compared to nonadaptive testing that were achieved with the MST design were a bit smaller (4–7%) but still of a relevant magnitude (Yamamoto et al. 2018b).

The next obvious step would be to allow for adaptation on the smallest possible level in order to harvest efficiency gains in the best possible way. For many ILSAs, this level would be testlets and, for some, even single items. Such a fine-grained adaptation would require the implementation of computerised adaptive testing (CAT; e.g. van der Linden and Glas 2000). Using CAT instead of MST seems to be an accessible option. Content balancing and item position effects (e.g. Debeer et al. 2014; Nagy et al. 2019) can well be accounted for in CAT, even though Yamamoto et al. (2018a) mentioned them as reasons to use MST instead of CAT. Both can be addressed in CAT (in a statistically optimal sense), for example, with the shadow-testing approach (van der Linden and Reese 1998). Of course, item position effects do not disappear by using CAT. However, controlling the position on which items are presented prevents systematic bias in ability estimates and group statistics.

Things are a bit different for the remaining major reason for favouring MST that is sometimes mentioned, namely, that response revision is possible within blocks or stages. This is not possible in typical testlet- or item-level CAT systems. However, several solutions for response revision in CAT have recently been proposed. Although some of these are very promising (e.g. Cui et al. 2018), the proposed methods have not yet reached operational status. In any case, it is worth developing these methods further because applying CAT in ILSAs makes further substantial improvements in measurement precision possible (e.g. Frey and Seitz 2011).

With regard to the operational aspects of conducting and maintaining ILSAs, substantial improvements can also be achieved through CAT. For example, adding new items to the item pool, identifying drifted items, and linking with previous assessments can be considerably simplified by adopting methods such as the continuous calibration strategy (CCS; Fink et al. 2018). The CCS incorporates all these aspects in the sense of a self-learning system and makes use of technology in order to optimise the functioning of test instruments *across* cycles, and thus widening the view from one cycle to seeing ILSAs as a process. Adopting such a view and the corresponding methods will make it easier to identify and to resolve issues that can be problematic for trend reporting. One example for such problems were the changes in the testing mode, the scaling method, and the type of some tasks incorporated in PISA 2015. The findings of Robitzsch et al. (2017) suggest that these changes could have biased the trend estimation for Germany. The CCS can help to introduce changes (such as new item types) more smoothly and more securely. The future success of ILSAs will largely depend on how well they adopt such methods to make optimal use of the enormous potential of computers. For computers to retain the status of essentially being used to imitate paper-based test procedures would hardly be viable for the future.

Measurement and Statistical Modelling of Context Variables

Traditionally, the main focus of ILSAs in education is on student achievement, and most of the assessment time is allotted to the assessment of *cognitive outcomes*, measured by means of achievement tests. In PISA 2015, for example, 120 min of assessment time were reserved for students to work on achievement tests, while they had 35 min to answer the student questionnaire. The variables assessed by questionnaire items (typically self-reports) are subsumed under *contextual information* (e.g. Organisation for Economic Co-operation and Development 2017a) or *context assessment* (Kuger and Klieme 2016). The amount of assessment time allotted to achievement tests compared to questionnaire variables stands in contrast to the number of constructs measured. While the 120 min of achievement testing in PISA 2015 were used to measure 10 literacy dimensions (reading, mathematics, and eight science scales), 32 derived variables (i.e. scales based on multiple items) in the student questionnaire were constructed from student questionnaire data collected in 35 min (Organisation for Economic Co-operation and Development 2017b). In addition to these derived variables, a number of student questionnaire variables are measured by individual items and, in addition to the student questionnaire, further context variables are assessed with context questionnaires for parents, teachers, and school principals.

The higher priority that is traditionally placed on achievement constructs is also visible in the methodology applied to achievement test data compared to questionnaire response data. The constructs assessed via achievement tests are modelled as latent variables with multidimensional IRT models. The multivariate distribution for the cognitive outcomes of the student population conditioned on a multitude of *background variables* from the context questionnaires is estimated using a latent linear regression model (Mislevy 1991). Finally, *plausible values* (PVs) are used as estimates of student proficiency. They are essentially multiple imputations of the cognitive outcomes for each student. In contrast, for constructs assessed via questionnaire items, observed scores or point estimates such as the weighted likelihood estimate are used in further analyses without conditioning.

In short, considerably more time, money, and analytical effort is invested in the measurement and modelling of cognitive outcomes than in that of context variables. However, the constructs assessed via questionnaires are receiving increasing attention. ILSAs use context variables not only as predictors of cognitive outcomes; several questionnaire variables are also treated as noncognitive educational outcomes (e.g. beliefs, motivation, and well-being) in their own right (e.g. Kuger and Klieme 2016). The increasing importance of context variables in ILSAs is visible, for instance, by the fact that the PISA 2003 and 2006 assessment frameworks contained a single page on "the context questionnaires and their use" (Organisation for Economic Co-operation and Development 2003, p. 18, and Organisation for Economic Co-operation and Development 2006, p. 14, respectively); in contrast, the frameworks for PISA 2009, 2012, and 2015 (Organisation for Economic Co-operation and Development 2009, 2013b, 2017a, respectively) have whole book chapters on the context questionnaires' framework.

The imbalance between the efforts invested in modelling cognitive outcomes versus context variables can be regarded as problematic for at least two reasons. First, using observed scores or point estimates for context variables as predictor variables in the population models means that measurement error for those constructs is not taken into account and data are assumed to be complete (i.e. not missing; Rutkowski and Rutkowski 2016). This is already a problem for the traditional focus on achievement using context variables merely as predictors, as missing data and measurement error can lead to biased estimates of the relationships between context variables and achievement (Rutkowski and Rutkowski 2016). Second, although the methods applied to achievement test data aim to produce unbiased estimates of multivariate population and subgroup distributions that have been corrected for measurement error, this is not achieved with the observed scores or point estimates for questionnaire variables. Consequently, the reported results of analyses for noncognitive outcomes (e.g. group means and standard deviations or regression coefficients) are affected by measurement error. Those on cognitive outcomes are not (or if they are, at least to a smaller extent), as the analysis of PVs provides estimates based on a latent regression model. That again implies that differences between subgroups within education systems will be underestimated for noncognitive outcomes. The practical implications of this difference can be illustrated by looking at brief reports such as "PISA 2015 results in focus" (Organisation for Economic Co-operation and Development 2018). The first major results table summarises achievement in science, reading, and mathematics by country; the second summarises results on students' science beliefs, engagement, and motivation. For both tables, significant differences between country means and the OECD average are highlighted. It is very likely that the results on noncognitive outcomes would contain more statistically significant differences and also larger effect sizes if the same elaborated modelling and estimation approach used to obtain the achievement results were applied to the context questionnaire data.

Another possible shortcoming of modelling the context questionnaire data, which also applies to the achievement data, is that the multilevel structure of the data (students in schools and/or classrooms) is not taken fully into account. The estimation of the population model in PISA 2015, for instance, was conducted with a linear regression model that did not incorporate random effects on the school level (Organisation for Economic Co-operation and Development 2017b). Depending on the strength of cluster effects, this could affect both the group means and standard errors obtained with the PVs based on the population model (Li et al. 2009).

To summarise, a methodological challenge for ILSAs is to reduce the imbalance between the modelling approaches used for achievement test data and context questionnaire data. Ideally, all constructs should be modelled as latent variables in order to account for measurement error and missing data, and the hierarchical structure should be included in the model. Practically, this is not yet feasible given the large number of constructs measured. Nevertheless, it is important to keep the ideal in mind and try to move towards it step by step.

Transparency of Data and Methods

In recent years, the *open science* movement has become increasingly influential in science, politics, and society, and public awareness of the importance of the accessibility, transparency, and replicability of scientific results has grown substantially. Standards that need to be met in order to achieve an open science culture include the citation of data and materials; the transparency and accessibility of data, methods, and materials; the preregistration of studies; and the replication of research findings (Nosek et al. 2015). For ILSAs, the aims of open science are particularly important as they are explicitly designed to provide policy makers with the information they need to make decisions on the configuration of education systems. Therefore, methods and results should be comprehensible and reproducible by independent experts. It has to be noted that ILSAs already meet some open science standards and have even been instrumental in setting some of these standards for the educational research community, particularly with respect to data and materials. The databases for ILSAs are made publicly available along with many of the instruments used in the assessment (except for the achievement test items for obvious reasons of test security and field trial data for most studies). Kuger et al. (2016) even made all translated versions of the PISA 2015 context questionnaires openly available. Furthermore, some of the open science standards are not particularly relevant for ILSAs. Standards related to preregistration are not applicable to ILSAs, as they are descriptive in nature and do not aim to test hypotheses specified a priori.

Nonetheless, there is room for improvement in ILSAs when it comes to meeting open science standards with respect to the transparency and accessibility of methods. Because ILSAs are often used as a reference by researchers working on smaller studies, it is particularly important that their methods are well documented and well founded. This is currently not always the case. For example, in PISA 2015, the root mean square deviation (RMSD) fit statistic was used to identify differential item functioning. RMSD values larger than 0.15 were interpreted to be indicators of deviations for achievement test items and values larger than 0.30 for context questionnaire items (Organisation for Economic Co-operation and Development 2017b). However, no rationale or evidence (e.g. based on simulation studies) was provided for these cut-off criteria. Another issue regarding documentation, which has been criticised by Rutkowski and Rutkowski (2016), is that the technical reports of several ILSAs are made available much later than the publication of the corresponding results, making it difficult for independent researchers to critically evaluate the methods of an ILSA in time. Finally, the transparency of methods should, in the case of statistical analyses, include the publication of the code used in the analyses (Nosek et al. 2015).

Most codes used for ILSA analyses have not been published in recent ILSAs, and proprietary software is often used that is not easily accessible to other researchers (e.g. the mdltm software [von Davier 2005] used in PISA 2015). A noteworthy and encouraging exception is the documentation of an Austrian national large-scale assessment, whose researchers published their methods as a comprehensive book

(Breit and Schreiner 2016) including the R code used for the analyses and in which the routines used for reporting were made available as an R package (Robitzsch and Oberwimmer 2018). In summary, there are possibilities for ILSAs to better meet open science standards, namely, the timely publication of technical documentations, the use of freely available software, and the publication of the analysis code used for the analyses and for processing the results.

Validation of Test Score Interpretations

According to the Standards for Educational and Psychological Testing, validity is considered "the most fundamental consideration in developing tests and evaluating tests" (American Educational Research Association, American Psychological Association,, and National Council on Measurement in Education 2014, p. 11). Measured against this central position of validity for assessment, the efforts made in ILSAs to provide evidence for the desired test score interpretations and the assumed effects on the future lives of the tested students are very sparse. This is surprising as several ILSAs formulate rather strong claims about the importance of the test scores for later life chances and individual success in society in general. As an example, the objective of PISA is defined as: "PISA assesses the extent to which 15-year-old students, near the end of their compulsory education, have acquired key knowledge and skills that are essential for full participation in modern societies." Empirical evidence is needed in order to support such test score interpretations.

Some such evidence was published based on the Youth in Transition Survey (YITS; Motte et al. 2008; Statistics Canada 2011) for Canada. YITS was set up to examine major transitions in young people's lives, with respect to education, training, and work. One cohort of YITS participated in PISA 2000 and a subsample of that cohort took the PISA reading test in 2009 again. The results show that PISA test results can indeed be connected to desirable outcomes. For example, a positive relationship between reading performance in PISA 2000 and the probability of attending a university at the age of 24 was reported (Organisation for Economic Co-operation and Development 2012b). Nevertheless, the study contained only a limited number of indictors for a *full participation in modern societies* and was conducted in Canada only. Another example for a study capable to provide some validity evidence is the TREE study (Transitions in Youth and Young Adulthood; Scharenberg et al. 2016). The study uses the Swiss PISA 2000 sample for several follow-ups. The results show, for example, that reading performance in PISA is predictive for educational attainment 10 years later. A substantial proportion of the students with low reading competency in PISA 2000 did not complete an upper secondary educational program (below proficiency level 1: 37%; proficiency level 1: 19%) compared to only 4% on each of the proficiency levels 4 and 5 (Scharenberg et al. 2016).

As the claims made by ILSAs are ambitious, more support from more countries is needed to justify the intended test score interpretations and uses in a broad sense.

Because the validity of test score interpretations and uses can change over time (Kane 2013), collecting validity evidence at one point in time only is not sufficient.

In order to justify the intended test score interpretations of ILSAs and the sometimes far-reaching decisions made by educational policy makers based on those interpretations, the future challenge is to implement longitudinal validation studies as an integral study part for all countries. Otherwise, ILSAs are based on unsupported claims; this does not do justice to the effort and costs invested, nor does it sufficiently support the conclusions derived.

Conclusions

This chapter outlined some of the challenges of ILSAs, as well as possible approaches on how to meet them. These relate to the following aspects: inclusion of new constructs, consideration of proficiency differences, measurement and statistical modelling of context variables, transparency of data and methods, and validation. The issues covered in this chapter are by no means exhaustive, e.g. we did not address challenges related to adaptation and translation of instruments and dealing with measurement invariance violations (e.g. Caro et al. 2014; Rutkowski and Rutkowski 2013). Due to the high methodological standards that have already been met, the ever-growing interest, the increasing use of data-based decision making, and the international character of ILSAs, it is likely that the discussed challenges can be met well.

ILSAs will probably continue to initiate and stimulate methodological developments in the future, to develop and examine these developments scientifically, and to apply successful approaches. Thus, they provide very good opportunities for addressing one of the Achilles heels of basic methodological-statistical research: mastering the leap from basic psychometric research to actual application.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*, 92–104.

Breit, S., & Schreiner, C. (Eds.) (2016). *Large-scale assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung* [Large-scale assessment with R: Methodological foundations of the Austrian educational standard evaluation]. Vienna: Facultas.

Caro, D. H., Sandoval-Hernández, A., & Lüdtke, O. (2014). Cultural, social, and economic capital constructs in international assessments: An evaluation using exploratory structural equation modeling. *School Effectiveness and School Improvement, 25*, 433–450.

Cui, Z., Liu, C., He, Y., & Chen, H. (2018). Evaluation of a new method for providing full review opportunities in computerized adaptive testing – Computerized adaptive testing with salt. *Journal of Educational Measurement, 55*, 582–594.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39*, 502–523.

Fink, A., Born, S., Frey, A., & Spoden, C. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling, 60*, 327–346.

Fraillon, J., Schulz, W., & Ainley, J. (2013). *International computer and information literacy study assessment framework*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

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.

Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2018). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*. Advance online publication.

International ICT Literacy Panel. (2002). *Digital transformation: A framework for ICT literacy*. Princeton: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/ICTREPORT.pdf

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Kaplan, D., & Kuger, S. (2016). The methodology of PISA: Past, present, and future. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning* (pp. 53–73). New York: Springer.

Kuger, S., & Klieme, E. (2016). Dimensions of context assessment. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning* (pp. 3–37). New York: Springer.

Kuger, S., Klieme, E., Jude, N., & Kaplan, D. (Eds.). (2016). *Assessing contexts of learning*. New York: Springer.

Li, D., Oranje, A., & Jiang, Y. (2009). On the estimation of hierarchical latent regression models for large-scale assessments. *Journal of Educational and Behavioral Statistics, 34*, 433–463.

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

Motte, A., Hanqing, Q., Zhang, Y., & Bussière, P. (2008). The youth in transition survey: Following Canadian youth through time. In R. Ross Finnie, R. E. Mueller, A. Sweetman, & A. Usher (Eds.), *Who goes? Who stays? What matters? Accessing and persisting in post-secondary education in Canada* (pp. 63–75). Montréal: Mc-Gill, Queen's University Press.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 assessment framework*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Nagy, G., Nagengast, B., Frey, A., Becker, M., & Rose, N. (2019). A multilevel study of position effects in PISA achievement tests: Student- and school-level predictors in the German tracked school system. *Assessment in Education: Principles, Policy & Practice*, *26*, 422–443.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture. *Science, 348*, 1422–1425.

Organisation for Economic Co-operation and Development. (2003). *The PISA 2003 assessment framework. mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2009). *PISA 2009 assessment framework. Key competencies in reading, mathematics and science.* Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2012a). *PISA 2009 technical report*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development. (2012b). *Learning beyond fifteen: Ten years after PISA*. Paris: OECD Publishing.
Organisation for Economic Co-operation and Development. (2013a). *Technical report of the survey of adult skills (PIAAC)*. Paris: OECD Publishing.
Organisation for Economic Co-operation and Development. (2013b). *PISA 2012 assessment and analytical framework mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.
Organisation for Economic Co-operation and Development. (2017a). *PISA 2015 assessment and analytical framework science, reading, mathematic, financial literacy and collaborative problem solving*. Paris: OECD Publishing.
Organisation for Economic Co-operation and Development. (2017b). *PISA 2015 technical report*. Paris: OECD Publishing.
Organisation for Economic Co-operation and Development (2018). *PISA 2015 results in focus*. https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf
Robitzsch, A., & Oberwimmer, K. (2018). *BIFIE survey: Tools for survey statistics in educational assessment. R package version 3.0–14.* https://CRAN.R-project.org/package=BIFIEsurvey
Robitzsch, A., Lüdtke, O., Köller, O., Kröhne, U., Goldhammer, F., & Heine, J. H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien. Eine Skalierung der deutschen PISA-Daten [Challenges in estimations of trends in large-scale assessments: A calibration of the German PISA data]. *Diagnostica, 63*, 148–165.
Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education, 8*, 259–278.
Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher, 45*, 252–257.
Rutkowski, D., Rutkowski, L., & Liaw, Y. L. (2018). Measuring widening proficiency differences in international assessments: Are current approaches enough? *Educational Measurement: Issues and Practice, 37*(4), 40–48.
Scharenberg, K., Hupka-Brunner, S., Meyer, T., & Bergman, M. M. (Eds.). (2016). *Transitions in youth and young adulthood: Results from the Swiss TREE panel study* (Vol. 2). Seismo: Zürich.
Singer, J. D., Braun, H. I., & Chudowsky, N. (2018). *International education assessments. Cautions, conundrums, and common sense*. Washington, DC: National Academy of Education.
Statistics Canada. (2011). *Youth in transition survey (YITS). Cohort A – 25-year-olds, cycle 6. User guide*. Ottawa: Statistics Canada.
van de Vijver, F. J. R., Jude, N., & Kuger, N. (in press). Challenges in international large-scale educational surveys. In B. Denman, L. E. Suter, & E. Smith (Eds.), *The SAGE Handbook of Comparative Studies in Education*. Thousand Oaks: Sage.
van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic Publishers.
van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259–270.
von Davier, M. (2005). *A general diagnostic model applied to language testing data (Research Report No. RR-05-16)*. Princeton: Educational Testing Service.
Yamamoto, K., Khorramdel, L., & Shin, H. J. (2018a). Introducing multistage adaptive testing into international large-scale assessments designs using the example of PIAAC. *Psychological Test and Assessment Modeling, 61*, 347–368.
Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018b). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement: Issues and Practice, 37*(4), 6–27.

Chapter 5
The Assessment Landscape in the United States: From Then to the Future

Eva L. Baker and Harold F. O'Neil Jr.

Assessment in Elementary and Secondary Schools

What is the state of assessment in US elementary and secondary schools? Certainly, in the current scene, times have changed from periods when testing occupied center stage of salient topics reported in media, and the foci of public policy, research, and practice. Why has testing contention faded from its historic centrality in US educational public policy discourse? Pointed questions have all but vanished about preemption of teachers' role by the use of external tests? Because in the past two decades, new teachers were hired into a test-driven accountability environment, many now know nothing different. Gareis (2017) points out that new teachers have also experienced test-based accountability from the vantage point of students. The day has been won by the advocates of testing with little if any substantiation that accountability provisions work, that is, that they have been shown to improve schools.

E. L. Baker (✉)
UCLA Graduate School of Education & Information Studies, Los Angeles, CA, USA

National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA, USA
e-mail: baker@cse.ucla.edu

H. F. O'Neil Jr.
University of Southern California, Los Angeles, CA, USA

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_5

Elementary and Secondary School Assessment

We focus on US Federal policy history leading up to the most current Federal provisions. Recall first that in the United States education is a function principally under the authority of each of the 50 states rather than a systematic Federal responsibility. However, the Federal government has considerable influence in that it can allocate marginal funds to states for particular programs with these funds providing disproportionate incentives.

US Federal Policy History on Assessment More than 50 years ago, the Federal government stepped in to the educational arena to address inequitable educational practices and lagging outcomes for poor and minority students. This disparity was verified by James Coleman in his landmark study of equality of educational opportunity (1966). Watershed legislation was passed by Congress in the first Elementary and Secondary Education Act (ESEA 1965) influenced as much by politics as research. This law notably introduced testing provisions to assure that underperforming and economically challenged students made demonstrated progress in academic performance and stands as the first major, national effort in accountability. This and subsequent reauthorizations of the law prescribed the use of commercialized standardized tests focused on general constructs such as reading and mathematics ability (see, for instance, the summary in Baker et al. 2016).

Almost 30 years later, in 1994, a significant revision of ESEA occurred, i.e., the Improving America's Schools Act (IASA 1994). IASA notably shifted in testing requirements from only disadvantaged students to include *all* students. IASA development had been strongly influenced by a series of high-profile, bipartisan reviews and recommendations: for example, *A Nation at Risk* (NCEE 1983) identified the reduction of US achievement; the *National Education Goals* (1999) articulated by the National Governor's Association specified goals in early childhood, literacy, teaching, and adult learning and including a fanciful goal that the United States would be "best in the world" in mathematics and science by the year 2000 (www2.ed.gov) and *Raising the Standards for American Education* (NCREST 1992). This letter review was issued by the National Council on Education Standards and Testing, a council of national and State legislators and measurement experts (the lead author participated).

In addition to broadening the focus from those economically deprived and minority students, IASA supported a set national standard to frame the assessments borrowing from practices in the United Kingdom. Standards-based assessments were supposed to be influenced by instruction, but in fact, it changed little of actual testing practice. Commercial assessments were marketed as standards-based, but in many cases on, relied on existing item pools and on traditional approaches to item development and psychometrics. Commercial publishers simply reformatted reporting from score averages to results expressed in terms of numbers of students reaching given thresholds (performance standards).

Alignment to curriculum and instruction was advocated, but took place *post* hoc rather than as in a coherent design (Baker 2005). IASA also provided for the use of noncognitive measures in accountability reports, but first ventures used archival information, such as school absences, rather than actual assessments of students' intrapersonal skills such as effort or self-efficacy.

ESEA legislation was again changed leading to the contentious "No Child Left Behind" (NCLB 2001). This legislation significantly modified IASA, with a mix of desirable and ultimately less useful provisions. For example, NCLB asked states for a plan to document Adequate Yearly Progress (AYP). These growth curves were to show how the yearly indicators of student achievement would eventually result in all students reaching the desired proficiency standards by 2014. Many states successfully gamed their design of AYP projections to start with small annual increments but sharp (and unrealistic) increases in later years in order to reach the 2014 proficient targets. AYP heightened attention to test-driven instruction. AYP results were to be disaggregated by subgroup to show disparities in performance to be overcome. If only one subgroup failed to make the AYP goal, the school was deemed failing. It isn't hard to estimate that more diverse schools, with more reportable subgroups, had increased probability of failure and these schools were often low-income schools.

On the positive side, NCLB included stronger equity provisions, requiring that a high percentage (95%) of students from identifiable groups, such as disadvantaged, English learning, participate in mandatory testing in order to combat inflated achievement results of schools and districts that encouraged poorer performing students to stay at home on testing dates. Like AYP, schools could be labeled "failing" if they did not meet participation levels. However, because most subgroups were each required to meet 95% participation, probabilistically more diverse schools were certain to fail over a 5-year period, just by chance. NCLB, like some of its predecessors, had consequences for failing schools. It required sanctions by states for districts and schools that persistently underperformed, after attempts at school improvement had been made.

Another set of education requirement later modifying NCLB was the Race to the Top legislation signed into law in 2009 as part of the American Recovery and Reinvestment Act (ARRA 2009), which was the Federal response to the worldwide financial downturn. The Race to the Top law created a competition among states for educational reform awards totaling more than $4 billion dollars. The RTT law included traditional ESEA provisions, such as developing and adopting common standards, requiring procedures for teacher and principal evaluation, supporting transition to standards, and developing longitudinal data systems based on individual performance. It also rewarded the development and expansion of charter schools.

The law prescribed, in addition to accountability and professional evaluation, the use of student data to improve instruction, a provision that implies different outcome measures. The absolute priority required for funding was that the state describes a comprehensive approach to reform where assessment was a strong component. Over three rounds of competition, 18 states (or 36%) were awarded amounts ranging from 500 million to 17 million dollars. However, the competition was

roundly criticized. EPI, for example, published a critique of a report by Shavelson et al., criticizing the use of student test scores to evaluate teachers and value-added analyses (Baker et al. 2010).

As part of the educational reform in this time period, and in recognition of the central role of assessment, the Federal government with ARRA funds held a competition for consortia of states to develop standards-based assessments to use in required standards-based assessment and to promote college and career readiness. These consortia were to prepare and to motivate others to develop more innovative approaches to testing. Two major awards were given, one to the Smarter Balanced Assessment Consortium (SBAC) and the other to Partnership for Assessment of Readiness for College and Careers (PARCC), each basing their assessments on the Common Core Standards in Math and English Language Arts (2010), projects that had been supported by private foundations and businesses. Starting with virtually all states participating, the consortia lost members over the next few years as states defected from the common core standards and chose less ambitious, less costly approaches. In any case, the innovation intended for the consortia has been marginal, perhaps because both developers and users were state policy makers already under various political and economic pressure. Another damper on innovation was that their contractors were drawn from the usual pool of standardized test purveyors.

Present-Day Federal Assessment Policy The most recent reauthorization of the ESEA is Every Student Succeeds Act (ESSA 2015). External testing for accountability unsurprisingly remains a part of required educational practice. The barebones of its provisions call for annual testing in reading and math in grades 3–8 plus once in secondary school. Science is to be tested at least once in each of the grade spans of 3–6, 6–9, and 10–12. Testing of English language proficiency is to occur annually in elementary and secondary grades for all those who qualify for English language development programs and to continue until students are transitioned to fluent English learner status. Assessment for students with special needs is also reflected in the ESSA law.

For the most part, assessments for all students are to be standards-based, but no longer emphasize common core standards (which fell into political disrepute). However, no criteria for assessment quality were provided. Some "flexibility" or variations in assessment are enabled in the legislation, for example, encouraging the development of innovative assessments and allowing the use of existing commercial tests at the high school level (operationally meaning the use of the SAT or ACT, college admission tests rather than measures of course based achievement). Graduation rates of 67% are required for each subgroup. Nonacademic measures are to be included in accountability reports, for example, student engagement, climate, safety, and postsecondary readiness. As before, archival information can be used, and the law now allows some process indicators such as the provision of resources, such as enrichment.

Parents are given the explicit ability to "opt out" of testing, but 95% rates of assessment participation are still required for each identifiable subgroup. However, in contrast to NCLB, falling below this percentage does not result in sanctions.

Moreover, expectations for making adequate yearly progress, the nature of interventions for underperforming schools, and reporting options are left to the devices of the various states. Disaggregating results by subgroup is still required, and new subgroups for reporting were added, including students with one or more parent in the military, homeless students, and students in foster care, the last three a chilling assessment of changes in American society. All told, ESSA represents a weakened form of accountability testing, and it is in large measure a compromised wrought by a Democratic administration and a Republican congress.

National Assessment of Educational Progress In addition to Federal statutes that influenced state assessment development, the Federal government itself manages additional accountability measures. Most important is the National Assessment of Educational Progress (NAEP), fully funded by the Federal government and administered by the by the National Center for Education Statistics (NCES), and is currently managed by the Educational Testing Service (ETS). NAEP also has a long history. Since the 1960s, NAEP has sought to provide overall and disaggregated estimates of the performance of students across the United States. It is administered on a sampling basis, with mathematics, reading, science, and writing domains regularly assessed for grades 4 and 8. There are aperiodic measures of other domains, such as Civics and Art. Although there is an extended trend line of performance reaching back several decades, for the last 20 years or so, each assessment has been generated using a domain-specific framework developed by educators and experts in assessment and contact domains. These frameworks and resulting assessments are intended to reflect more contemporary views of content and skills as currently taught in schools. Results from the NAEP administrations typically offered on a 2-year cycle, are provided overall, disaggregated by subgroups, such as gender, race/ethnicity, disabilities, proxies for socioeconomic level, and identified English learners. Although at its inception NAEP's purpose was to lead in assessment innovation, that goal has had only sporadic attention recently, focused on technology options.

Originally, great pains were taken so that NAEP would not be used to compare state and local jurisdictions and lead to undue focus on the test content during instruction. Instead, as advocates of accountability grew more insistent, NAEP changed its sampling and reporting protocols so that results could be compared not only on broad geographic bases and across administration cycles but to a more explicit comparative framework. NAEP changed from gross regional reporting, e.g., West, South to reports of state by state achievement. Later, 25 large city schools were also sampled sufficiently to provide usable data so that their achievement could be compared with one another as well with states.

Across the states, from cycle to cycle, there is variation in achievement, with high-performing states having both high levels of achievement on NAEP and high graduation rates. But the interpretation of NAEP data is conflicted. For example, from 1960 to present, the overall trajectory of NAEP is largely flat, meaning that no major improvement has been found. However, there has been discernible growth in some aspects of NAEP achievement. Over last 20 years, there has been more

noticeable at particular percentile levels, e.g., 25th. There is also variation in US states. The best performing states have between 80% and 87% of students scoring at or above the basic level (the lowest standard) and between 40% and 53% scoring at or above proficient in 4th grade math.

It should be noted that some measurement experts have had difficulty with NAEP achievement levels (or performance standards) thinking that they were set too high. There have been sunnier interpretations of overall data; for instance, we have seen noticeable improvement, in closing the gap between Latino and White students. But overall, not much has happened. Particularly disappointing given the federal and state resources invested in education during this timeframe.

There are numerous explanations for the NAEP data, some more apologetic than others. Here is a brief reprise of them: (1) The frameworks guiding NAEP development have only loose connections with state and district curriculum emphases, so that instruction is similarly loosely connected. At best, NAEP could be considered as measures of transfer from school learning. (2) Students taking NAEP have little incentive to perform well, as the test doesn't "count" in any noticeable way. Without incentives, performance can lag, especially when there is a surfeit of testing for students. (3) The educational system has absorbed during this time period many new types of students, with myriad background and language issues, all of which could impact NAEP performance. (4) The lower socioeconomic level of US society has dropped below the wealthiest sectors. Students are coming from struggling families who may have less opportunity to help their children learn. (5) Technology may not be supporting learning. Screen time is up, with 2017 data suggesting that students between 8- and 10-year-old spent an average of 6 h a day, up about 50% from 2015 and those 11–14 in front of screens an average of 9 h a day. With time split among mobile devices, computers, and television, much of the screen time is spent in social interaction, such as texting not much time for reading is left. (6) The numbers for out of school reading are not easily compared, but overall, the United States is one of the lower countries in number of minutes read a day, and the results diminish with age. So, it is likely that children are not seeing much modelling of intellectual activity at home. (7) NAEP is not high stakes for students, teachers, or principals; thus there is little teaching to the test. So, in combination with other findings, it is reasonable to propose that the NAEP findings actually represent US achievement.

Assessment in Workforce

Workforce readiness is a concern of educators at the K-12 and postsecondary levels. It also has a long history. For example, in 1991, Secretary's Commission on Achieving Necessary Skills (SCANS) identified twentieth century (now twenty-first-century skills desired by employers). These involved key cognitive demands, personal responsibility, and interpersonal skills. Although there is a history of specifying workforce requirements to inform schooling, two major innovations

should be noted. First, workforce expectations entered specifically into the K-12 curriculum as 20th and cognitive readiness (O'Neil et al. 2014). These skills included domain independent skills like problem-solving and communication, personal behaviors like timeliness, and readiness to learn and interpersonal skills like teamwork (Pellegrino and Hilton 2012). Personal requirements like abstaining from drug use are also considered. These general topics seemed to recur in periodic surveys of employers (NACE 2014). In higher education, annual surveys have reported on the perceptions of college students about their job readiness. Only about 4 in 10 feel well prepared for their careers after graduation and feel unprepared in transitional skills, like resume writing and interview preparation, as well as areas like problem-solving. Gender gaps also occur with men feeling better prepared than women despite that data favor women in graduation rates and access to graduate and professional education.

Trends like job insecurity and wage stagnation provide incentives for students and institutions to focus more heavily on career preparation rather than foundational skills, which are usually the focus of college and university. For liberal arts, a recent study (Jaschik 2016) reported almost 9% loss of major in the humanities, and universities regularly report starting salaries of graduates by their major course of study. The large average burdens of student loans (currently just under $40,000 inhibit university students accepting low-paying, entry-level jobs https://studentloanhero.com/student-loan-debt-statistics).

One interesting resource is the analysis of needed competencies that is widely available. The US Department of Labor has sponsored website (O∗Net) that provides a list of potential occupations supported with each illustrated by the set of skills required. This site also shares updates on the potential availability of occupations for job seekers of all ages. We expect an increase in assessments relevant to workforce skills at both the pre-collegiate and higher education levels.

An Innovative Workforce Assessment System: Training Assessment Framework (TAF)

We now briefly report the CRESST design and findings of a multiyear Navy assessment project that applies assessment lessons learned from the K-12 and higher education settings, as well as insights from the science of assessment. A new system for assessment is under development by CRESST for the Navy Education and Training Command (NETC), the group responsible in the Navy for preparing young enlisted US sailors for more 70 career paths.

It is based on previous CRESST models for assessment design (Bewley et al. 2009) implementation and validation. The model has a number of important attributes. First, it is designed to serve multiple purposes, including certification of end-of-course competencies. Second, the system will support aggregation of performance and achievement results across disparate courses and jobs. A third

purpose was to provide feedback to instructors and curriculum designers. A fourth and central purpose of the project was to provide an additional indicator to validate the job classification decision made at the point of recruitment. The decision is based predominantly on the ASVAB (https://www.military.com/join-armed-forces/asvab).

The project team will develop reports (both graphical and in text) tailored to different audiences and commit to develop innovative technical approaches to validate inferences for the full range of purposes. The Training Assessment Framework design document (Baker and Choi 2018) lists user needs, cognitive demands, domain knowledge acquisition, and task specifications to delimit parameters to guide assessment development. These parameters will allow future of automated design systems. This framework also address validity, e.g., development comparisons of expert and novice performance.

Two courses for two different Navy jobs or ratings were the proof of concept for the Training Assessment Framework Model to determine whether the process was generalizable. The very different Navy jobs were damage control (ship protection) and fire control (a radar-technician-like job). The former involved dangerous situations, team environments, and problem-solving and procedural knowledge. The second is a technical electronics job that involved operations and maintenance skills. Navy subject matter expects in these areas helped determine relevant content for assessment. This step was followed by extensive development of ontologies documenting content for each rating (Baker and Choi 2018).

Three major types of formats comprised the examination. The first was the generally familiar selected response format that used multiple choice, drag and drop, and hotspots to show answers. The second involved performance assessments using scenario-based simulations asking for search and procedural knowledge on integrated tasks likely to be confronted by the sailor on the job. The third was a knowledge map (O'Neil and Chung 2011) that asked sailors to create nodes and links to reveal them understand of the hierarchy, structural and functional relationships among important elements of the job knowledge domain. The maps are scored against expert maps. Trainees were also asked to complete affective scales measuring domain-specific self-efficacy and anxiety so as to allow us to understand the degree to which different test formats influenced affective responses.

We were constrained to use a technology platform (iPad) and to limit testing time. The examinations were administered on a pre-and post-instructional basis to trainees' subject matter experts (SMEs) to determine a benchmark for validity. The report of this work (Baker and Choi 2018) documents the design options, describes the data, and presents innovative psychometric and validity solutions to performance and knowledge mapping items.

In the future, the project will scale-up with additional Navy jobs. Our commitment to develop assessments with a heavy emphasis on their technical quality (e.g., validity evidence) will have relevant implications for civilian workforce assessment and for K-12 systems as well.

Critical Discussion

The accretion of requirements for K-12 assessment in federal law, for the most part, has been decisive in the management of schools. Because of the historical concern with accountability, terms framing instruction as "test driven" or "evidence based" are frequent, with almost no discussion of the quality of the evidence. Earlier warning about the limitations of instruction geared to what could be measured seems to have evaporated as a major concern. Alignment studies may address standards but more attention is given mapping instruction to actual test specifications. Despite professional and academic attention to appropriate requirements for design, development, and validity, there is little evidence that validity is taken seriously, despite widespread recognition of the Standards for Validity (AERA, APA, NCME 2014, 1999, 1985). For instance, the notion of policy capture, that is, what experts think performance should be related to standard setting for achievement, has little in the way of empirical validity, and it is an essential part of criterion-referenced reporting models (Baker 2012). Moreover, there is rare validity attention to the full range of purposes for tests, e.g., certification, accountability, or instructional improvement, and none is easily accessed on the public-facing websites by testing groups.

To end this tour, let's consider some trends in society that are influencing assessment. The first, which we have informally observed, is there is less attention and reliance on technical quality indicators derived from high-quality research. Moreover there is a general retreat from scientific bases of quality mirrors everyday life, where stars, likes, and other populist indicators of quality have seemingly replaced reliance on technical expertise and evidence. Although most commercial tests received some level of analysis, many focus solely on content validity; that is, are important content and skills measured and are they representative of desired content? A second general purpose motivated by the legal system is the verification of fairness and equity; that is, are all student given a fair opportunity to succeed?

Second, psychometrics has continued to evolve and indicators will move toward findings that can be directly interpretable for improvement of learning, as opposed to those procedures that require transformations that take findings somewhat far afield from their original design. One positive development has been the exploration of feature analysis as a validity and design check on assessment. Here cognitive demands, task requirements, and domain content are qualitatively rated on existing or newly developed items, and then relationships of item features and performance on the test are examined (Baker and Cai 2014; Baker 2015; Madni et al. 2018).

Finally, there is the impact of technology as a pervasive element in modern life, no less in testing. Technology has been used in ways to make assessment more efficient (computer-adapted testing) and more palatable, in game and simulation-based tests, in automating scoring (Burstein 2003) and in supporting automated design. The energy of the startup community in testing, with their race to market, has once again served to reduce commitment to careful validation studies, seen by the startup community as not to add much value and slows access to the market.

Unless we can continue to develop infrastructure tools, such as automated ontology extraction, automated scoring and reporting, and simulated students for accelerated data collection, we fear that the demand for data and the propensity to develop empirical findings to drawn validity inferences will die a relatively quick death. However, we look enthusiastically to the future where we believe assessment, technical quality, and technology when properly combined (Baker et al. 2016) will continue to drive greater utility, motivation, consequences, and innovation in assessment and learning.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

American Recovery and Reinvestment Act of 2009, Pub. L. No. 111-5.

Baker, E. L. (2005). Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era* (pp. 315–335). Mahwah: Erlbaum.

Baker, E. L. (2012). Standards for educational and psychological testing. In J. A. Banks (Ed.), *Encyclopedia of diversity in education* (Vol. 4, pp. 2076–2081). Thousand Oaks: SAGE.

Baker, E. L. (2015, May). *Feature analysis: Improving the validity of competency tests.* Presentation at the colloquium on assessment and accountability implications of competency and personalized learning systems. Boulder, CO.

Baker, E., & Cai, L. (2014, August). *CRESST analysis of the quality of a state assessment* (Technical slide report). Los Angeles: CRESST.

Baker, E. L., & Choi, K. C. (2018). *Training Assessment Framework* (Report to funder). Los Angeles: University of California, Los Angeles, National Center on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Barton, P. E., Darling-Hammond, D., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010, August). *Problems with the use of student test scores to evaluate teachers* (Briefing Paper #278). Washington, DC: Economic Policy Institute. Available at http://epi.3cdn.net/724cd9a1eb91c40ff0_hwm6iij90.pdf

Baker, E. L., Chung, G., & Cai, L. (2016). Assessment gaze, refraction, and blur: The course of achievement testing in the last hundred years. *Review of Research in Education* (Centennial Issue), *40*, 94–142.

Bewley, W. L., Chung, G. K. W. K., Delacruz, G. C., & Baker, E. L. (2009). Assessment models and tools for virtual environment training. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 300–313). Westport: Praeger Security International.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross disciplinary approach* (pp. 113–121). Mahwah: Lawrence Erlbaum Associates.

Coleman, J. S. (1966). *Equality of educational opportunity*. Oxford: U.S. Department of Health, Education.

Elementary and Secondary Education Act of 1965 as amended, 20 U.S.C. §241 (1974).

Every Student Succeeds Act of 2015, Pub. L. No. 114-95.

Gareis, C. R. (2017). *Assessment leadership: Leveraging performance-based assessments for deeper learning*. Presented at the 21st Annual School-University Research Network Leadership Conference at the College of William & Mary, Williamsburg VA.

Improving America's Schools Act of 1994, Pub. L. No. 108, Stat. 3518.

Jaschik, S. (2016, March). The shrinking humanities major. *Inside Higher Ed*. https://www.insidehighered.com/news/2016/03/14/study-shows-87-decline-humanities-bachelors-degrees-2-years

Madni, A., Kao, J. C., Rivera, N. M., Baker, E. L., & Cai, L. (2018). *Exploring career-readiness features in high school test items through cognitive laboratory interviews* (CRESST Report 857). Los Angeles: University of California.

National Association of Colleges and Employers (NACE). (2014). *NACE 2013–2014 career services benchmark survey for colleges and universities*. Bethlehem: Author.

National Commission on Excellence in Education. (1983). *Nation at risk*. Washington, DC: Department of Education.

National Council on Education Standards and Testing. (1992, January 24). *Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American People*. Washington, DC: Government Printing Office.

National Education Goals Panel. (1999). *The national education goals report: Building a nation of learners, 1999*. Washington, DC: U.S. Government Printing Office.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, §115, Stat. 1425 (2002).

O'Neil, H. F., & Chung, G. K. W. K. (2011, April). *Use of knowledge mapping in computer-based assessment*. In Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

O'Neil, H. F., Perez, R. S., & Baker, E. L. (Eds.). (2014). *Teaching and measuring cognitive readiness*. New York: Springer.

Pellegrino, J. W., & Hilton, M. L. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: The National Academies Press.

Part II

Chapter 6
Monitoring Student Achievement in Austria: Implementation, Results and Political Reactions

Claudia Schreiner, Birgit Suchań, and Silvia Salchegger

Introduction

The Education System in Austria

The Austrian education system is hierarchically organised, highly centralised and one of the few systems that is selective at a very early age. The creation and implementation of pre-primary legislation (kindergartens and crèches) is undertaken by each province (Austria consists of nine *Bundeslaender*). The Federal Ministry for Education, Science and Research (BMBWF), the federation, is primarily responsible for school legislation; however, school governance is divided between the federation and the provinces. The third column of the education system, the tertiary level (universities, universities of applied sciences and university colleges of teacher education), is also currently managed by the BMBWF, under its capacity science and research.

Compulsory education lasts for 9 years beginning at age six. Additionally, all children must attend kindergarten for at least 1 year before starting primary school. This obligatory year in kindergarten is a relatively new element in pre-primary education. Introduced in 2010, it is one outcome of discussions of the results from system monitoring activities in Austria. An important aspect of the Austrian school system is early selection into two different educational tracks following primary school. At the age of ten, children and their parents must choose between lower-level secondary *general* school (*Neue Mittelschule – NMS; Grades 5 to 8*) or

C. Schreiner (✉)
Department of Teacher Education and School Research, University of Innsbruck, Innsbruck, Austria
e-mail: claudia.schreiner@uibk.ac.at

B. Suchań · S. Salchegger
Federal Institute for Education Research, Innovation and Development of the Austrian School System (BIFIE), Salzburg, Austria

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_6

lower-level secondary *academic* school (*Allgemeinbildende Höhere Schule – AHS; Grades 5 to 12*). Four years later, at the age of 14, they must again decide on their educational path. At this stage the Austrian school system is characterised by a wide range of education and training tracks, including an extensive vocational sector in which 80% of 15- to 19-year-olds are enrolled (Eurydice 2018).

Developing and Implementing Achievement Monitoring

Austria's experience with international large-scale assessments dates back to the Computers in Education Study (COMPED) in 1992 and the Third International Mathematics and Science Study (TIMSS; all populations) in 1995. Even though both studies were barely perceived by the media and the public, the results of TIMSS led to the establishment of a support system for teachers called IMST, which still exists.[1] The introduction of PISA in 2000 (the Programme for International Student Assessment) marks the beginning of the regular collection of student achievement data in Austria. The PISA 2000 results were above the OECD average and received little media attention. The subsequent drop in student's achievement in PISA 2003 (due to weighting issues within vocational apprenticeship schools in PISA 2000) lead to increased policy debate and paved the way for participation in further studies (Eder and Altrichter 2009). As a consequence, Austria also participated in the international assessments TIMSS (2007 and 2011; Grade 4 only) and PIRLS (2006, 2011 and 2016).

Public debate on education was furthered by attention to the results from international studies after the first two PISA studies in all German-speaking countries. This was a main driver for the shift from input-based regulation towards a stronger focus on outputs. As well as participating in additional international assessments, national efforts addressed the possible merits of introducing national educational standards, and researchers discussed ways of defining educational standards and corresponding assessment systems (e.g. see Klieme et al. 2003).

In Austria, the discussion progressed from a strong idea of accountability and minimum achievement standards towards a development-oriented approach. Educational standards and a national assessment system were introduced based on competencies in the sense of a broader quality of a person (Weinert 2001) and competency models. Educational standards, assessments and the feedback system were intended to stimulate systematic and target-oriented quality development at school and classroom level by means of data-oriented self-regulation processes and measures considering contextual conditions. As such, educational standards were introduced in a legal act in 2008/2009. The educational standards describe in the form of can-do statements what students should be able to do in the subjects German and

[1] For more information about IMST, see https://www.imst.ac.at/texte/index/bereich_id:8/seite_id:8

mathematics at the end of primary school (Grade 4) and for German, mathematics and English as a foreign language at the end of lower secondary (Grade 8). The regular assessment of national standards was introduced in the school year 2011/2012 and is now conducted as a census of one grade per year in one subject. One year the assessment takes place for mathematics in Grade 8, the next in mathematics in Grade 4, then in English in Grade 8 and so on. In order to assess all grade-subject combinations, a 5-year cycle is needed. All schools represented by their school leaders and teachers receive feedback on their students' learning outcomes. Student's performance on these tests has no influence on their grades or school career. Students are granted access to their results as a sign of appreciation for their participation and to help establish a sound self-concept.

This development – including reinforcing the participation in international assessments as well as the introduction of national assessments – led to the foundation of the Federal Institute for Educational Research, Innovation, and Development of the Austrian School System (BIFIE) and its legal anchoring in the year 2008. This institute was established as a publicly financed but independent organisation that would conduct studies in applied educational research in order to monitor the school system and provide information for evidence-oriented quality development from school level up to educational policy.

International and National Assessments Today

The combination of national and international assessments yields a comprehensive system of data and feedback on different levels (from the student up to the system as a whole; see Fig. 6.1). Although the various levels of operation (from system level to the level of schools, teachers and an individual student) have different needs

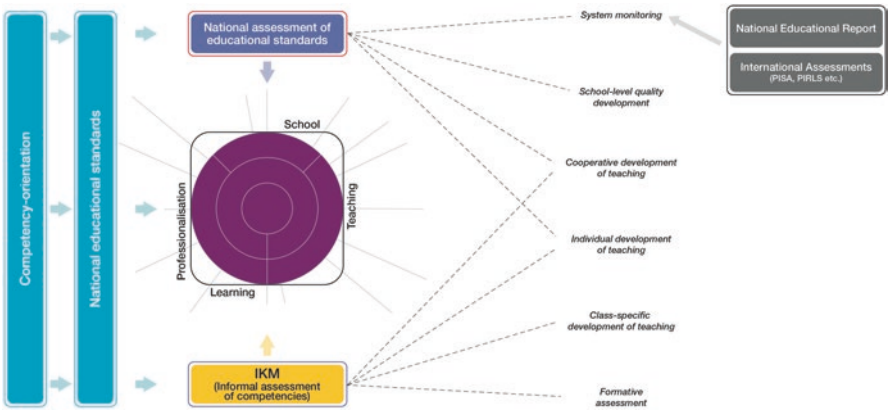

Fig. 6.1 National and international assessments in Austria and their intended impact on school quality

and modes of action, they share the basic idea of evidence-oriented quality development (see below).

The national standards-based assessments provide feedback at all levels, to teachers, for developing their teaching, to schools and the system level (both nationally and regionally). The assessment of educational standards takes place at the end of primary and lower secondary school and therefore sheds light on students' learning outcomes in a summative perspective at points of intersection in the educational system. The assessment is conducted as a census of one subject and one grade each year and comprises a comprehensive system of individualised feedback, including local school authorities, school leaders, teachers and students (with their parents). Nationwide results as well as the outcomes at the province level are published in reports. Researchers can use the anonymised data from the national assessments via a research database (see https://www.bifie.at/bildungsforschung/forschungsdatenbibliothek/).

The system level information provided by the national assessments is enhanced by the broader perspective that the international assessments provide by their international context and comparison. Currently, Austria participates in TIMSS, PIRLS and PISA as international student assessment programmes. The BIFIE acts as national study centre for Austria. With regard to the previously described school system, Austria collects international data at the end of primary school, with PIRLS and TIMSS, and again towards the end of compulsory schooling when students attend a wide range of programmes.

Tools for assessing student achievement by teachers in form of a self-assessment (IKM – informal assessment of competencies) complement the system. The self-assessment tools are linked to the national education standards but were designed as a platform with assessment packages for teachers to gather feedback on their students' competencies at earlier points (Grades 3, 6 and 7). These tools are therefore instruments for teachers to adapt their approach for the class and for formative assessment.

All of these assessments are publicly financed. Schools (and teachers) have the responsibility to work with the data and feedback they received from national assessments; however, no accountability system is in place regarding the results.

The regular delivery of data from international and national large-scale assessments reignited the discussion on the central issues and problems of the educational system, such as social inequality. The National Education Report largely contributed towards focusing discussions on the school system. This report, established in 2009, is organised by the BIFIE in a 3-year cycle. It comprises one volume with indicators from different sources of data and one volume taking up several educational key topics. The authors represent a wide range of the educational research community in Austria. The National Education Report provides a comprehensive fundament for educational policy debate as well as steering processes within the school system.

Key Findings of International and National Assessments in Austria

Two findings from international and national assessments elicited a lot of media-related, political and scientific discussion in Austria: (1) Austria has a relatively high rate of low-performing students who do not reach the baseline proficiency level, and (2) the association between social background and academic achievement is relatively high, with students from advantaged homes performing, on average, much better than those from disadvantaged homes. In the following we present more detailed findings on both aspects.

High Rate of Low-Performing Students

According to the Austrian constitution, no young person should leave school without having acquired basic skills: "Each young person should […] be enabled of participating in the cultural and economic life of Austria, Europe, and the world" (Austrian Federal Constitutional Law, Article 14). Consequently, the school system should be designed in order to keep the ratio of low performers small and to provide as many students as possible with the basic skills necessary to enable them to actively participate in cultural, economic and social life.

The amount of low-performing students in Austria has been estimated by various assessments each defining "low achiever" in its own way (see Table 6.1, for an overview). In the international assessments PIRLS and TIMSS, low achievers are defined as those students who remain below International PIRLS or TIMSS Benchmark 2. In PISA students who remain below PISA baseline Proficiency Level 2 are regarded as low achievers. Students performing below this PISA baseline level are considered to be at serious risk of not being sufficiently enabled to participate effectively and productively in life (Organisation for Economic Cooperation and Development 2016a). Austria's national educational standards define quite high standards to aim for regarding "what students should normally be able to do". The national assessment reports are based on different competency levels, where Level 2 refers to students who reach the targets defined in the national standards.[2] This includes being able to flexibly use acquired knowledge and skills in the respective subjects and areas. Students who do not reach this goal are divided into two groups: the ones that partially reach this goal by showing the ability to fulfil routine tasks without the flexible use required at Level 2 ("Level 1") and students who do not reach the national standards and also have problems fulfilling routine tasks ("below Level 1"). Students performing below Level 1 are considered to be at considerable

[2] Only in the domain of English, proficiency levels as well as the national curriculum are based on the CEFR *(Common European Framework of Reference for Languages)*.

Table 6.1 Share of low-performing students in Austria compared to the EU average

Share of low-performing students in Austria compared to the EU average																	
				Share of low-performing students in reading (%)						Share of low performing students in mathematics (%)					
Assessment	Grade	Number of participating EU countries	Definition of "low-performing students"	AUT		EU		Difference AUT-EU		AUT		EU		Difference AUT-EU	
TIMSS 2011	4	21	Students below TIMSS International Benchmark 2	–		–		–		29.57	(1.75)	26.36	(0.30)	3.20	(1.78)
PIRLS 2016	4	24	Students below PIRLS International Benchmark 2	15.62	(1.13)	17.99	(0.22)	**−2.37**	(1.15)	–		–		–	
PISA 2015	Mainly 10	25[a]	Students below PISA Proficiency Level 2	22.54	(1.04)	20.15	(0.21)	**2.39**	(1.06)	21.75	(1.08)	22.31	(0.22)	−0.56	(1.10)
BIST-UE M4 2013	4	1	Students who do not or only partially reach the goals defined by the national education standards (below national level 2)	–		–		–		23.10	(0.00)	–		–	
BIST-UE M8 2017	8	1		–		–		–		42.60	(0.00)	–		–	
BIST-UE D4 2015	4	1		38.49	(0.00)	–		–		–		–		–	
BIST-UE D8 2016	8	1		44.58	(0.00)	–		–		–		–		–	

Notes: Significant differences ($p < 0.05$) are in bold; standard errors are in parentheses
Source: Own computations based on the respective datasets
EU European Union, *TIMSS* Trends in International Mathematics and Science Study, *PIRLS* Progress in International Reading Literacy Study, *PISA* Programme for International Student Assessment, *BIST-UE* National Standards Based Assessment, *M* Mathematics, *D* German
[a]Only countries which participated in the computer-based assessment are included

risk of their lack of basic competencies hindering their further personal and educational development.

As shown in Table 6.1, PISA 2015 identified 23% of low-performing students in reading and 22% of students performing below baseline Level 2 in mathematics. For Grade 4, the international assessments PIRLS 2016 and TIMSS 2011 revealed 30% of low performers in mathematics and 16% in reading. Moreover, the national standards based assessments (each based on a census of either fourth or eighth graders) revealed that about 40% of Austrian students fail to demonstrate the reading skills required by the rather strict educational standards, which are based on the national curriculum (see Table 6.1). In mathematics, this share is 23% at fourth grade and 43% at eighth grade.

Table 6.1. moreover shows that Austria's share of low performers in mathematics is similar to the EU average. In reading, it is significantly lower than the EU average at fourth grade but significantly higher for 15-year-olds. These results are disappointing in the light of the fact that Austria is among the EU countries with the highest education spending (Organisation for Economic Cooperation and Development 2018, Indicator C1). Altogether these results indicate a lack of effectivity in Austria's education system.

More detailed analyses show that student achievement (including the number of low-performing students) varies decisively by students' backgrounds. These results are discussed in the following section.

Low Equity in Education

Programmes such as PISA seek equity in much the same way as outlined in the Austrian constitution, where "[one basic aim of schooling is] to secure the highest possible education level for the whole population, *independent of origin, social situation, and financial background* […]" (Austrian Federal Constitutional Law, Article 14). According to the Organisation for Economic Cooperation and Development (2013, p. 13), PISA defines equity in education as "providing all students, regardless of gender, family background or socio-economic status, with opportunities to benefit from education". Moreover, PISA stresses that although "equity does not imply that everyone should have the same results. It does mean, however, that students' socio-economic status or the fact that they have an immigrant background has little or no impact on their performance, and that all students, regardless of their background, are offered access to quality educational resources and opportunities to learn" (Organisation for Economic Cooperation and Development 2013, p. 13). As such, equity requires that the level of education attained (or performance achieved) does not depend on students' background characteristics, such as their parents' socio-economic status or immigration history.

Detailed analyses highlighted the following findings:

- In Austria 25% of the variance in reading performance of 15- to 16-year-olds can be explained by family background. Therefore, the association between social

background and performance in Austria is one of the highest among OECD/EU countries (Oberwimmer et al. 2016, p. 178).

- The academic performance of immigrants is on average 2 years behind that of Austrian students after 9 years of schooling. The difference between the groups can not only be attributed to their immigrant background but also to substantial differences in socio-economic status (Vogtenhuber et al. 2012, p. 154; Herzog-Punzenberger et al. 2012, p. 11; Salchegger et al. 2016).
- The academic performance of children of parents showing lower levels of educational attainment (i.e. parents who have finished compulsory schooling at most) is on average 3 to 4 years of schooling behind that of children whose parents have finished some kind of tertiary education (Oberwimmer et al. 2016, p. 181). Moreover about 70% of children from parents showing low educational attainment do not or only partially reach the nationally defined standards, whereas a decisively lower proportion (about 20%) of children with highly educated parents do not reach the nationally defined standards in mathematics (Oberwimmer et al. 2016, p. 183).
- A disproportionate number of youths with immigrant background comprise at risk students: In reading, 18% of native students perform below Proficiency Level 2, but 39% of immigrant students perform below Level 2 in PISA 2015 (Salchegger et al. 2016).

These results show that there are large differences in educational achievement between social groups and between immigrants and natives at the end of compulsory schooling.

Austria's early tracking has been discussed in this light. As mentioned above, the first selection of students into different school tracks takes place after Grade 4, when students are usually 10 years old. Students need to have attained (very) good grades in primary school to be formally permitted to attend an *academic* track school *(Allgemeinbildende Höhere Schule)*. However, these (very) good students (respectively their parents) can also choose a *general* track school *(Neue Mittelschule)*. All other students (i.e. students whose grades do not meet the criteria for academic track attendance) attend a general track school. In this way Austria is, next to Germany, the country with the earliest selection of students into different school tracks across the Organisation for Economic Cooperation and Development (see Organisation for Economic Cooperation and Development 2016b, p. 167).

Analyses of the attendance rates of academic track schools have shown that students with low-level educated parents attend academic track schools far less frequently than students with highly educated parents (Bruneforth et al. 2016). However, this difference can only partially be explained by differences in previous achievement. Effectively, students from lower social origin have a double disadvantage: Not only do children with a more advantageous social origin on average perform better from the beginning of their educational career (primary effect of social origin), equally able students from families with a high social status more often attend the *academic* track (secondary effect of social origin). Of fourth grade students who performed on the Austrian average in mathematics, only 24% of those

whose parents have a low level of formal schooling attend an academic track school after Grade 4 compared to 64% of those whose parents finished tertiary education. Therefore, a child's educational pathway is greatly determined by their family background – independent of actual achievement (see also Bruneforth and Itzlinger-Bruneforth 2015).

Attending an academic track school is connected not only with a higher probability of university attendance but also with the phenomenon that a "cluster" of high-performing students (with high-level educated parents) creates an environment more conducive to learning. As a consequence, attending an academic track school can enhance the probability of completing higher education in various ways.

How to Go Further?

Changes of paradigm have taken place in two different aspects since the political and public sectors became aware of the results of international large-scale assessments. First, public and political attention was directed towards outputs and outcomes of the school system. Thus, one change of paradigm concerns the shift from input-driven to output-oriented policy making. Second, public and political discussion was influenced by the facts and data presented in the studies. This concerns the paradigm of evidence-based policy, which is more ambiguous as the first.

Whereas the quantity and quality of public debate has definitely improved, basing decisions and reforms on the evidence of large-scale assessments has turned out to be extremely challenging. Regardless, international assessments have highly influenced educational debate, and the following two reforms, at least, were driven by discussions on assessment results:

1. The implementation of the national educational standards in 2008/2009 and their regular assessment as of 2011/2012. This formally marks the shift in policy towards a focus on outputs and making it transparent that the required outputs comprise being able to use knowledge and skills. On the one hand, the educational standards complemented the international assessments. On the other hand, the feedback system for schools and teachers broadens the scope of evidence-oriented quality development. As national competency standards define the learning goals much more clearly than the curriculum, they serve as orientation for teachers, students and parents.
2. The centralisation of the matriculation examination (*Matura*), which provides access to universities. This reform is also a landmark in the Austrian education system. It was adopted in 2009 and implemented beginning in 2014. Whereas the school leaving exams at the ends of Grades 12 and 13, respectively, were previously the responsibility of the subject area teachers of each school, the new *Matura* consists of a centrally administered written examination conducted on the same date for all students as well as non-standardised assessments related to the specific focus of the school.

However, the debate on measures to improve equity in education is much more challenging than the above-described and already implemented reforms. Since this is one of the main (recurring) results in international and national assessments and because it has been discussed in conjunction with a significant feature of the Austrian school system, the early tracking at the age of 10 years, we would like to outline the political and reform debate that took place in this context. Lassnigg and Vogtenhuber (2015, p. 18) characterised this debate in the following way: "… on the one hand there is wide consensus that the choice is too early, but as the academic track is sacrosanct for broad and influential groups on the other hand, there is no solution for the early choice". The current – and altogether third – attempt (following initiatives already in 1922 as well as between 1970 and 1986) towards a comprehensive school for 10- to 14-year-olds was initiated in 2007/2008 as a pilot project and anchored in law in 2012. This attempt resulted in the establishment of a new school type called New Secondary School (*Neue Mittelschule – NMS*), replacing all *general* secondary schools since 2015/2016. Actually, NMS was originally meant to be a comprehensive school for all students aged 10–14 years (replacing both general and academic secondary schools) in order to avoid early tracking and the consequences on equity in education. Because this initial plan was unenforceable and the academic track (AHS) still exists alongside the NMS, the early tracking is still in place in Austria. Thus, the third attempt to modify the transition from primary school to lower-level secondary school resulted (once again) in a reform trying to improve the secondary general track instead of a realisation of a comprehensive school system.

The discussion of the early tracking system and the resulting new school-type NMS is only one (significant) example of how national and international assessments influence public debate in the educational sector in Austria. It also shows how difficult it is in fact to implement far-reaching but necessary reforms on the system level based on evidence. Various factors, many of which have very little foundation in evidence, influence political decision making as well as enforcing and implementing reforms. However, by fuelling public discussion in various fields, national and international large-scale assessments have at least indirectly led to various changes within the educational landscape in Austria. The following list gives an exemplary overview of major reforms:

- One year of compulsory kindergarten before primary school since 2010 (already mentioned above).
- New legislation on the employment of teachers (2013) and new teacher education system (2015) to provide a common standard for teacher education.
- Establishment of a "school entry phase" in 2016 (including the last year of kindergarten and the first 2 years in primary school) in order to improve the transition between pre-primary education and primary schools.
- Compulsory education and training until the age of 18 after finishing compulsory schooling. The aim of this reform (implemented since 2016) is to broaden the competencies of disadvantaged young people and to reduce the percentage of out-of-school population in Austria in youth, which is one of the highest across

all OECD countries at the age of 15–16 years (Organisation for Economic Cooperation and Development 2016a, p. 290, Table A2.1).
- School autonomy package (2017) to reorganise responsibility regarding the organisation of school time and learning groups as well as staffing recruitment.

Coming back to the model presented above, the data collected in national and international studies (in a reliable and objective way) are supposed to serve as a basis for quality development. However, finding the way from data to concrete actions has proven to be another great challenge in the Austrian education system in many contexts. This is particularly evident regarding schools and teachers, who receive feedback from national assessments. Research on the use of data feedback in quality development processes show that data per se does not lead to action designed to improve the work of the school or life in the school.

At the beginning of the 2000s, Helmke suggested a three-phase process for the optimal use of data in school quality development (Helmke 2004). This involves *reception*, reading the feedback; a phase of *reflection*, discussing the school's results and conditions, incorporating knowledge about the school from other sources, putting the data in context and thereby reaching a deeper understanding of the data's meaning; and third, taking *action*, reacting to the feedback the data provide and proactively deciding how to tackle future challenges. Evidence underpins these three phases of response for school leaders as well as teachers (Schreiner and Wiesner 2016; Wiesner et al. 2018). Moreover, structural equating shows that the reflection phase is a strong mediating factor between reception and taking action, with no direct path from reception to action (Wiesner et al. 2018). In regard to system monitoring in Germany Maritzen (2014) reached a similar conclusion: "There is no direct way from measuring to shaping, neither in politics nor in regards of teaching" (p. 406). As such, jumping from the reception phase directly to action (skipping the reflection phase) has been shown not to work. Putting the data feedback into context, reflecting on what the data mean for one's school also including a discussion on the visions, goals and conditions, is essential so as to be able to decide on measures and provisions for the future. This is a collaborative process, which is typically cyclical and in reality probably seldom as simple and linear as the model suggests.

How this process works can be shown for schools striving to enhance quality in teaching, learning and living in the school and also teachers working on their way of teaching. In principle, this also applies for quality development on a system level. However, the process becomes even more complex at the level of policy making due to the complexity of the system concerned.

Contextualising data to create meaningful evidence for driving action is a very challenging process.

Nevertheless, a path of evidence-oriented quality development for all levels of the school system has been introduced in a very consistent way. This in fact is remarkable in the Austrian context and regarding the dominant traditions in the system. Now in place for 6 years, the assessment and feedback system is the current norm for school leaders and schools. Despite this extensive and successfully imple-

mented measure of national standards-based assessments, a political debate on changing the assessment system to strongly focus on the individual student and their pathway in the school system has arisen in the past two years based on the new government programme. It remains to be seen how this debate evolves and whether scientifically sound measures on student achievement will further underpin quality development on the system and school level in Austria.

References

Bruneforth, M., & Itzlinger-Bruneforth, U. (2015). Die Schulwahl von Schüler/innen am Ende der 8. Schulstufe im Lichte ihrer Mathematikkompetenzen. In M. Stock, P. Schlögl, K. Schmid, & D. Moser (Eds.), *Kompetent – wofür? Life Skills – Beruflichkeit – Persönlichkeitsbildung. Beiträge zur Berufsbildungsforschung* (pp. 263–282). Studienverlag: Innsbruck.

Bruneforth, M., Vogtenhuber, S., Lassnigg, L., Oberwimmer, K., Gumpoldsberger, H., Feyerer, E., … Herzog-Punzenberger, B. (2016). Indikatoren C: Prozessfaktoren. In M. Bruneforth, L. Lassnigg, S. Vogtenhuber, C. Schreiner, & S. Breit (Eds.), *Nationaler Bildungsbericht Österreich 2015, Band 1: Das Schulsystem im Spiegel von Daten und Indikatoren* (pp. 71–128). Graz: Leykam. https://doi.org/10.17888/nbb2015-1-C

Eder, F., & Altrichter, H. (2009). Qualitätsentwicklung und Qualitätssicherung im österreichischen Schulwesen: Bilanz aus 15 Jahren Diskussion und Entwicklungsperspektiven für die Zukunft. In W. Specht (Ed.), *Nationaler Bildungsbericht. Österreich 2009. Band 2: Fokussierte Analysen bildungspolitischer Schwerpunktthemen* (pp. 305–322). Graz: Leykam. https://doi.org/10.17888/nbb2009-2-C1

Eurydice. (2018). *National education systems. Austria overview.* Retrieved from https://eacea.ec.europa.eu/national-policies/eurydice/content/austria_en

Helmke, A. (2004). Von der Evaluation zur Innovation: Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. *SEMINAR – Lehrerbildung und Schule, 2*, 90–112.

Herzog-Punzenberger, B., Bruneforth, M., & Lassnigg, L. (2012). *National education report Austria 2012*. In *Indicators and topics: An overview*. Leykam: Graz.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Berlin: BMBF.

Lassnigg, L., & Vogtenhuber. S. (2015). *Challenges in Austrian educational governance revisited. Re-thinking the basic structures* (IHS Sociological Series Working Paper 107). Retrieved from http://irihs.ihs.ac.at/3586/

Maritzen, N. (2014). Glanz und Elend der KMK-Strategie zum Bildungsmonitoring. Versuch einer Bilanz und eines Ausblicks. *Die Deutsche Schule, 106*(4), 398–413.

Oberwimmer, K., Bruneforth, M., Siegle, T., Vogtenhuber, S., Lassnigg, L., Schmich, J., … Trenkwalder, K. (2016). Indikatoren D: Output – Ergebnisse des Schulsystems. In M. Bruneforth, L. Lassnigg, S. Vogtenhuber, C. Schreiner, & S. Breit (Eds.), *Nationaler Bildungsbericht Österreich 2015, Band 1: Das Schulsystem im Spiegel von Daten und Indikatoren* (pp. 129–194). Graz: Leykam. https://doi.org/10.17888/nbb2015-1-D

Organisation for Economic Cooperation and Development. (2013). *PISA 2012 Results: Excellence through equity: Giving every student the chance to succeed (Volume II)*. Paris: OECD Publishing. https://doi.org/10.1787/9789264201132-en.

Organisation for Economic Cooperation and Development. (2016a). *PISA 2015 results (Volume I): Excellence and equity in education*. Paris: OECD Publishing. https://doi.org/10.1787/9789264266490-en.

Organisation for Economic Cooperation and Development. (2016b). *PISA 2015 results (Volume II): Policies and practices for successful schools*. Paris: OECD Publishing. https://doi.org/10.1787/9789264267510-en.

Organisation for Economic Cooperation and Development. (2018). *Education at a glance 2018: OECD indicators*. Paris: OECD Publishing. https://doi.org/10.1787/eag-2018-en.

Salchegger, S., Wallner-Paschon, C., Schmich, J., & Höller, I. (2016). Kompetenzentwicklung im Kontext individueller, schulischer und familiärer Faktoren. In B. Suchań, & S. Breit (Eds.), *PISA 2015. Grundkompetenzen am Ende der Pflichtschulzeit im internationalen Vergleich* (pp. 77–100). Graz: Leykam.

Schreiner, C., & Wiesner, C. (2016, June). *Wandel von Routinen als Voraussetzung für Veränderung. Wissenstransferanalysen mit Blick auf Bildungsstandards, Kompetenzorientierung und datenbasierte Rückmeldungen.* Speech presented at the 22. EMSE conference. Retrieved from https://www.researchgate.net/publication/306142646/download

Vogtenhuber, S., Lassnigg, L., Gumpoldsberger, H., Schwantner, U., Suchań, B., Bruneforth, M., … Eder, F. (2012). Indikatoren D: Output – Ergebnisse des Schulsystems. In M. Bruneforth, & L. Lassnigg (Eds.), *Nationaler Bildungsbericht Österreich 2012. Band 1. Das Schulsystem im Spiegel von Daten und Indikatoren* (pp. 111–164). Graz: Leykam.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (pp. 17–31). Weinheim: Beltz.

Wiesner, C., Schreiner, C., & Breit, S. (2018, September). *Reflectioning and proflection in schools: What are schools really doing with data?* Speech presented at the ECER Conference.

Chapter 7
Use of Assessments to Inform Educational Policies in French-Speaking Belgium

Dominique Lafontaine

In French-speaking Belgium, the national assessments developed only lately, and there are not yet national assessments developed by professionals that can be used to evaluate trends. Therefore, the only tools available to rigorously evaluate trends are international assessments. French-speaking Belgium has participated in international assessments since the early 1970s and their results are highly valued by policy-makers. Their level of awareness of the strengths and weaknesses of the education system can be considered as good. However, until recently, the impact of international assessments on education policies has been limited. Only scattered initiatives have been taken. From 2014, an extremely ambitious plan called *Pacte pour un enseignement d'excellence* has been launched. The Pact tackles most of the systemic weaknesses of the education system in FS Belgium and addresses at the same time structural change (lowering grade repetition, moving to a comprehensive lower secondary education), curricular changes and governance in a long-term perspective. A significant reform of the system of national assessments is currently under discussion.

Introduction

Belgium is a federal state. Besides the federal state, there are three "communities", defined on the basis of language: the Flemish-, French- and German-speaking communities. Since 1989, the three education systems and their policies have been managed autonomously by their respective Ministries of Education. National assessment policies are different in each community and even the decision to participate in international assessments is taken at the community level. In view of this, covering

D. Lafontaine (✉)
University of Liège, Liège, Belgium
e-mail: dlafontaine@uliege.be

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_7

the three different systems in one contribution is not practicable. The focus will therefore be on the French community, for the simple reason that the author of this contribution is from the French community and is more knowledgeable about educational policies and national and international assessments in this community. When the term "national" is used, it refers to French-speaking (FS) Belgium, not to Belgium as a nation.

The education system in Belgium is a non-comprehensive one. The orientation towards academic or vocational tracks officially starts at the end of grade 8. Beyond that point, the system tends to stream pupils according to their abilities. The proportion of pupils attending special education is rather high (4% of all pupils) (Fédération Wallonie-Bruxelles 2016). Rates of grade repetition are extremely high, the highest of any OECD country: by the age of 15, nearly one student out of two (46%) has repeated at least one grade (Quittre et al. 2018). There is no catchment area: students and their families can freely choose the school they want to attend without any geographic limitation. This "free choice" results in huge competition between schools, which develop strategies to attract students. Middle-class parents develop strategies in order to find the "best" school for their children. This has been described by sociologists as a "quasi-market" (Dupriez and Maroy 2003). It results in huge differences in school intake: on the one hand, "sanctuary" schools attended mostly by middle-class students; on the other hand, "ghetto" schools attended by lower-class students or students from a migration background. The system is highly segregated from both an academic and a sociocultural point of view (Monseur and Lafontaine 2009; Monseur and Lafontaine 2012). Since the end of the 1980s (1988), more resources have been allocated to schools attended by a high proportion of students coming from underprivileged backgrounds ("positive discrimination policy") (Demeuse and Monseur 1999; Friant et al. 2008).

Teacher training for primary and lower secondary education takes the form of a 3-year bachelor's programme and is organised in teacher training institutes (not in university faculties of education). For upper secondary education, teachers are trained at university, in a consecutive training model. They first study a subject (e.g. mathematics, science or French language) for at least 4 years, and then an additional half or full year is dedicated to pedagogy, educational psychology, didactics and practical training in the field.

Up to 2006, there was no external certificate assessment or examination at any level. All examinations were internal, designed by the teacher for his/her own students. Schools issued diplomas on the basis of their own criteria, and obviously the level of standards and value of diplomas were not equivalent from school to school. External "diagnostic" assessments had been set up in 1995, focusing on teachers and aimed at providing them with useful information about their students' strengths and weaknesses in three domains: reading, mathematics and science. This kind of assessment, which still takes place, has no impact at all on students' careers, which is why it will not be described in detail here (for additional information, see http://www.enseignement.be/index.php?page=25162&navi=2024). Since 2006, external certificate assessments have been gradually adopted (in accordance with the Law of June 2006), firstly at the end of primary education (certificate of basic education,

grade 6), secondly at the end of lower secondary (grade 8, in mathematics, mother tongue, science and foreign languages only) and thirdly at the end of upper secondary (grade 12). At that level, the scope of the assessment is very narrow (history in the academic track, reading informational texts in the vocational track), meaning that the certificate of upper secondary education is still delivered mostly on the basis of internal and local standards. The external certificate assessments are developed under the responsibility of the Compulsory Education Monitoring Service of the Administration of the Ministry of Education (http://www.enseignement.be/index.php?page=26245). The developers are teachers, supervised by inspectors. They have plenty of good will, very limited resources, and limited professional knowledge about assessments, cognitive item development and basic notions of psychometrics. A limited field trial is organised, but the psychometric properties of the tests are not controlled for.

Finally, and this is important with respect to international assessments, the system of national external assessments has been designed so that all tests are publicly available as soon as they have been taken. Items are obviously under strict embargo up to the time of testing; after that, all the material is released, so that parents, teachers and pupils can use the previous versions of the tests for information and training. Consequently, as no item is kept under embargo, items cannot be reused from one year to the next for comparison and calibration purposes. Although the developers try to assess the same knowledge and skills each year, variations in the results cannot be interpreted as reflecting changes in students' achievement. In other words, in terms of monitoring of the education system, there is no rigorously designed national assessment (such as NAEP in the USA or tests in the Netherlands developed by the CITO), supervised by assessment professionals, that can be used to evaluate trends. The only tools available to rigorously evaluate trends are international assessments. This has not prevented the media and others from using the national assessments as evidence of "true" variations in students' achievement, even though academics have repeatedly pointed out their limitations.

The History of International Assessments

The French community has regularly taken part in the IEA studies (Six Subject Study, SIMSS, IEA Reading Literacy 1991, TIMSS 1995) (Lafontaine and Blondin 2004); since the beginning, results were reported separately for the Flemish and French communities. Despite alarming outcomes in FS Belgium, especially in science and reading comprehension, the impact on education policies – if any – has been very small. The IEARL results in 1991 (Lafontaine 1996) and the TIMSS results in 1995 (Monseur 1997) were reported in the media (to a lesser extent than PISA, however), but had limited or no impact on educational policies.

The power of the OECD and its eagerness to disseminate the PISA message as widely as possible have obviously led to the PISA results being more visible in the French community than the IEA studies. In 2001 especially, but also at the time of

the following PISA cycles, many articles/features were published in newspapers and magazines (even the most popular ones, such as TV magazines), debates on radio and TV channels were organised and PISA experts were regularly asked to give lectures to educational audiences (such as schools, teachers' unions, inspectors, principals and parents' associations) and public audiences (such as political parties). Since December 2001, every sneeze in the educational field has been related – rightly or wrongly – to PISA! This has been PISA's achievement: when it comes to educational matters, it has become the measure of everything, a kind of all-purpose assessment, supposedly capable of providing answers to all questions.

International and/or National Assessments Today

Since 2000, the French Community of Belgium has participated in PISA, in PIRLS (2006, 2011 and 2016) and most recently in TALIS (2018). As the example of PISA will be examined hereafter, it is important to say a few words about the results in PIRLS and their dissemination. In the three cycles in which FS Belgium has participated, it has ranked last among EU and OECD education systems. National reports have described the situation and given recommendations in terms of curriculum and teachers' initial and in-service training (Lafontaine et al. 2017). Meetings and discussions have been organised with the Ministry of Education, as well as with the Monitoring Compulsory Education Board. The decision has been taken by the authorities to avoid holding a press conference, on the grounds that the disastrous results would undermine teachers' morale. In other words, when OECD pressure is not there to push decision-makers to disseminate the results widely, the impact of international assessments is similar to what it was in the 1980s and the 1990s. Knowledge of the results is restricted to a handful of researchers and stakeholders; decision-makers in education play the politics of the ostrich.

As a representative of FS Belgium in the IEA General Assembly, I have several times highlighted the interest and value of participating in other international studies. Despite the authorities' interest in participating, the main argument put forward for not doing so was a financial one. While it is true that budgets allocated to international assessments are limited, they could obviously be increased if the political will were there.

An Example in Question: PISA Results and Their Links with Educational Policies

An overview of the PISA results and trends is shown below with minimal use of figures. Those interested in more detail can access the "national" reports on the page http://www.aspe.ulg.ac.be/ or consult Baye et al. 2009; Demonty et al. 2013;

Lafontaine et al. 2003; Quittre et al. 2018; Lafontaine et al. 2019. Three main topics will be discussed: achievement, equity and segregation. As a reminder, three domains (reading, mathematics and science) are assessed in each cycle, but with a major focus on one domain.

Achievement

Since 2000, performances and, more broadly speaking, all indicators in the FS community have been remarkably stable (Quittre et al. 2018) (Table 7.1). For the sake of clarity, we will comment only on the results of the cycles in which each domain is the major domain. In mathematics, the mean performance in all cycles was close to the OECD average: 498 in 2003, 493 in 2012 and 489 in 2015. In science, the mean performance is also stable across the different cycles, but significantly below the OECD average (486 in 2006 and 485 in 2015). In reading literacy, the mean performance was well below the OECD average from 2000 to 2006. In 2009 and 2012, the mean performance increased and reached the OECD average (501); however, in 2015, it decreased again (483). There is no room here to develop an interpretation of this positive (followed by a negative) trend. According to our analyses (Lafontaine 2014), the change does not result from change in the curriculum or teaching practices, because no curricular change was implemented. Instead, it seems to be related to a structural reform in the organisation of lower secondary education (grades 7 and 8). In 2006–2007, in parallel with the introduction of a certificate assessment at the end of primary education, a reform ("réforme du 1er degré") introduced remedial classes for students who had failed this examination. Since then, the

Table 7.1 Overview of the PISA results in French-speaking Belgium

PISA	2000	2003	2006	2009	2012	2015
Performance in reading	**476**	476	473	**490**	501	483
Standard deviation	**96**	101	105	**94**	99	101
Percentage of low-performing students (below level 2) in reading	**28**	25	26	**23**	19	23
Percentage of high-performing students (levels 5 and 6) in reading	**7**	7	7	**10**	10	6
Performance in mathematics		**498**	490	488	**493**	489
Standard deviation		**108**	109	104	**96**	92
Percentage of low-performing students in mathematics		**23**	24	26	**24**	24
Percentage of high-performing students in mathematics		**16**	14	12	**12**	10
Performance in science			**486**	482	487	**485**
Standard deviation			**103**	108	97	**96**
Percentage of low-performing students in science			**24**	25	21	**23**
Percentage of high-performing students in science			**7**	6	5	**5**

Note. The major domain per cycle is shown in bold. The time series begins when the domain is the major domain for the first time: 2000 for reading, 2003 for mathematics, 2006 for science

programme for these low achievers has been adapted: they receive extra classes, especially in mother tongue and mathematics, the goal being to prepare them to take this examination again until they pass it. Previously, students were oriented towards pre-vocational classes with no obligation to gain their certificate of basic education (CEB). It seems that this reform (since abandoned), combining the effects of an external assessment, a clear standard to reach and additional courses in the two main domains, led not only to an increase in the mean performance in reading, but also to a quite substantial reduction of the proportion of low achievers in reading (9% decreasing between 2000 and 2012).

Apart from this temporary improvement in reading performance in 2009 and 2012, the PISA results are congruent with the results of IEA surveys: TIMSS 1995 and PIRLS 2006 to 2016. Students' performances, especially in reading literacy and science, have been – and still are – a matter of concern (Lafontaine et al. 2017; Monseur 1997). These quite poor performances in the three domains (which are, at best, close to the OECD average) correspond to an uneven distribution of the proportion of students of different levels of ability, especially in reading: FS Belgium has fewer high achievers (levels 5 and 6 on the PISA scale) than OECD countries on average, but above all has a higher proportion of low achievers.

In reading in 2000 (Lafontaine et al. 2003), for instance:

- 7.5% of the students showed an excellent level of reading literacy (level 5 and above). This proportion was only somewhat lower than the OECD average (9.5%).
- 28% of the students scored below level 2 (very low reading ability). This proportion of low performers was much higher than the average in the OECD countries (18%).

In mathematics, this asymmetry does not apply: the proportion of low and top performers is very close to the OECD average. In science (2006), as in reading, there were somewhat fewer top performers (7%) than the average in OECD countries (9%), but significantly more low achievers (24% in FS Belgium and 19% on average).

The most striking result in the French community is the huge gap between top and low performers. The French community's education system has one of the highest such gaps (standard deviation), although it has been decreasing over time. The width of the dispersion is likely to be related to the frequent use of grade repetition in the French community. The French community has the highest rate of grade repetition among OECD countries: in PISA 2015, 46% of 15-year-olds had repeated at least one grade, and 13% out of this 46% had repeated at least two grades. In addition, differences of achievement between academic and vocational tracks are very large (68 points on the PISA scale between academic and vocational tracks at grade 10) (Quittre et al. 2018).

Socioeconomic Gap

Since the beginning of PISA, numerous analyses have focused on equity, namely, the relationship between students' sociocultural background (parents' socio-professional status, parents' level of education, migration background, family wealth and resources, cultural communication and so on) and achievement in reading, mathematical and scientific literacy (OECD 2011, 2016).

In terms of equity, the French Community of Belgium appears as one of the education systems in which the gap in performance according to socioeconomic status is the highest (Fig. 7.1.). In all cycles, the gap between the top quartile of students from the most privileged backgrounds and the bottom of most underprivileged students is more than 100, while the OECD average is around 80 (88 in 2015) (Quittre et al. 2018). Students who are vulnerable in terms of socioeconomic background are more at risk of being among the low achievers in the French Community of Belgium than similar subgroups of students in the majority of other education systems.

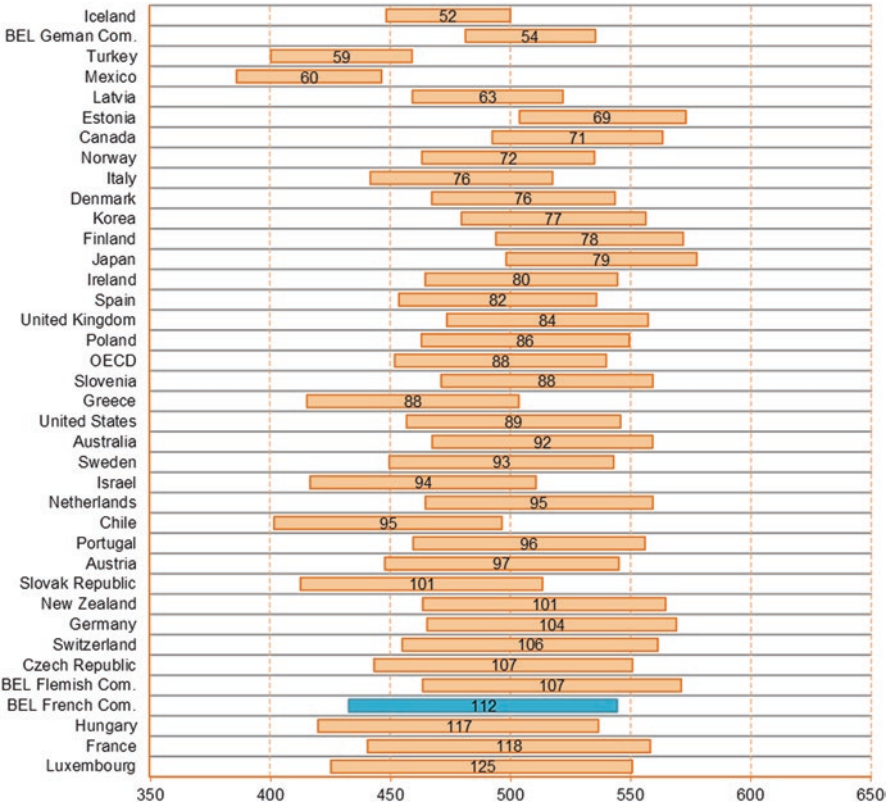

Fig. 7.1 Gap according to the index of students' socioeconomic status. PISA 2015 data

Segregation

In French-speaking Belgium, the between-school variation is high: the performances in the PISA test are very different from one school to another. By comparison, in the Nordic countries, school performances are very similar. The situation in FS Belgium has slightly improved over time: in 2006, the between-school variance was 46%; in 2015 it decreased to 42% (Quittre et al. 2018). As shown in Fig. 7.2, the percentage of the differences between schools explained by the socioeconomic composition of the schools is very high, one of the highest among OECD countries (around 75%). This means that students are sorted not only according to their abilities, but also according to their social background. One of the major lessons learned from PISA is that the French community education system is not successful in coping with or compensating for social inequalities. Although equity is highly valued in

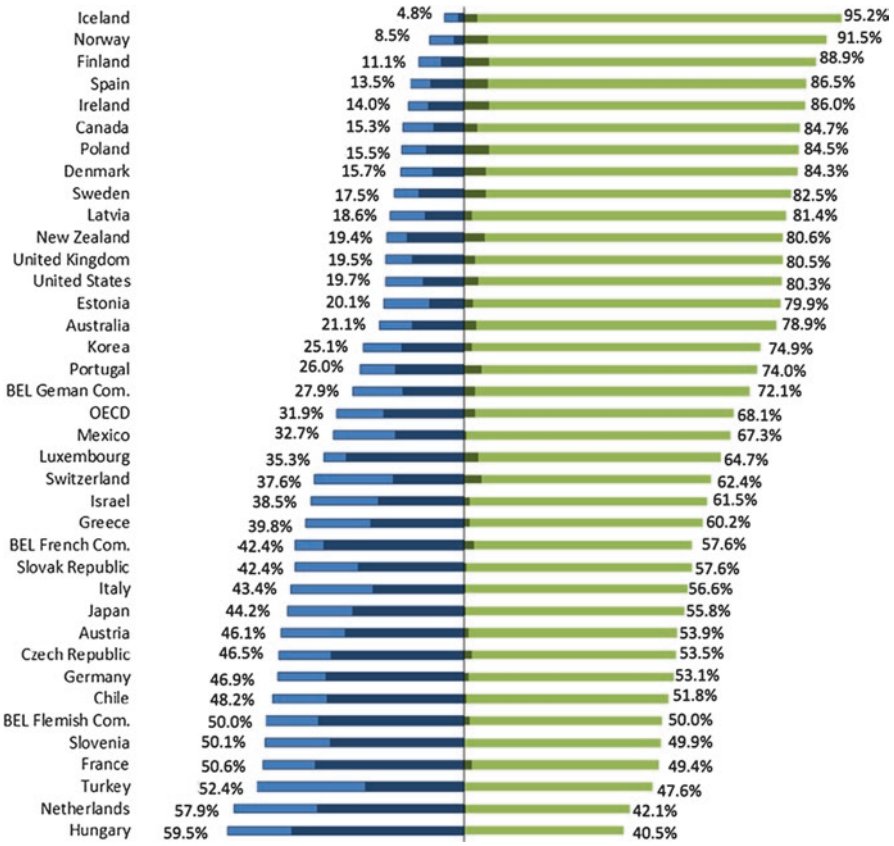

Fig. 7.2 Breakdown of the variance between schools and within schools and proportion of the variance explained by the socioeconomic status of the students and of the school (in bold). PISA 2015 data

official laws (especially in the "Décret Missions", in which the goals of the education systems are defined), and despite compensating policies (positive discrimination), the education system remains highly segregated, as demonstrated by PISA in an unequivocal way.

This overview of the results was the key message delivered by the successive national reports and by the PISA experts through reports, lectures, and interviews about PISA. The relative underachievement in reading and science calls for changes in teaching practices or curricular measures, while low equity and high school segregation call for more fundamental thinking about the core structure of the education system – the non-comprehensive or streaming/"segregated" approach reflected in grade repetition, early assignment to tracks and complete freedom of choice of schools (resulting in the so-called quasi-market).

How to Go Further? Critical Discussion

Before going into details, we want to underline that the results of PISA and other international assessments are valued by decision-makers in FS Belgium. PISA is neither strongly criticised nor disregarded. PISA and the results of other international studies are congruent with other indicators. The annual publication (Fédération Wallonie-Bruxelles 2016) issued by the Compulsory Education Monitoring Service also shows alarming results in terms of grade repetition, dropout and percentage of students leaving the education system at 18 without any certificate from secondary education.

PISA experts such as Prof. D. Lafontaine and her team are influential and often consulted by decision-makers; every Minister of Education since 2000, regardless of political affiliation, has paid careful attention to the PISA results. However, although efforts have been made to provide information to schools (principals and teachers), it cannot be taken for granted that PISA is well known by these categories of actors. Nevertheless, the main stakeholders represented at the Board of Compulsory Education – inspectors, teachers' unions, parents' associations and heads of education networks (organising state schools, municipalities and provinces, Catholic schools) – are well aware of the main PISA results. In terms of information and awareness among stakeholders, the situation is quite positive. In other words, the lack of major reform cannot be attributed to a lack of awareness.

Between PISA 2000 and 2015: Change in Educational Policies

No Change in Curricula

In terms of curricula, no change has been implemented since 2000. In the late 1990s, new standards were adopted defining key competencies for primary and secondary education, and most stakeholders hoped that these new standards and a

competency-based pedagogy would contribute to an enhancement of students' knowledge and skills. In the field of reading, there were scattered initiatives, but no major or comprehensive action was taken to change teachers' practices, despite the fact that the national reports, especially the PIRLS ones, clearly showed that teachers' practices for reading literacy in FS Belgium were outdated and that pupils definitely had fewer opportunities to learn effective reading comprehension strategies than in more effective education systems (Lafontaine et al. 2017; Lafontaine et al. 2018).

Change in Terms of Governance

One domain in which PISA might have acted as a catalyst is the governance of the education system. Discussions around the creation of a Monitoring Compulsory Education Board and external assessments had started before the release of the PISA 2000 reports, but the process may have been sped up by PISA. A law adopted by Parliament in 2002 established a Board in which all the stakeholders involved in education are represented: the administration, the school organising authorities, teachers' unions, inspectors, researchers in education, teacher training colleges and parents' associations. The goals of this Monitoring Board include:

- Guiding educational reforms and helping their implementation
- Developing a coherent system of indicators
- Defining priorities for in-service teacher training
- Organising external assessments in order to improve the quality of education

Equity and Segregation: Structural Reforms

After the results of the first PISA cycles, discussions and debates around equity issues and segregation were numerous. Many policy-makers and politicians, especially those on the left wing, saw the inequity revealed by PISA as unacceptable.

Between 2003 and 2009, the two governments in which the Ministers of Education were socialist defined a plan called the "Strategic plan for education" in which several initiatives to reduce inequity and segregation were defined:

- Strengthening the resources allocated to underprivileged schools (Friant et al. 2008).
- Regulating the enrolment of pupils in secondary education (grade 7) and increasing the social mix in schools. Several versions of the "Décret Inscriptions" were adopted between 2007 and 2010. Despite the fact that this attempt to regulate the enrolment was restricted to grade 7 and preserved families' right to choose a school, this law met with massive opposition from privileged families and the principals of the most privileged schools.

Apart from that, up to 2015, no major initiative was taken regarding the structural organisation of the education system, namely, the rate of grade repetition and early tracking. Several studies, including OECD reports, had highlighted the link between these stratification features and lack of equity (Mons 2007; Monseur and Lafontaine 2009; Monseur and Lafontaine 2012; OECD 2011, 2016). Between 2009 and 2014, a government formed from the same parties (socialists and Christian democrats), but in which the Minister of Education was a Christian democrat, explicitly opted for a "no-more-top-down-reform" approach in education; the Minister deliberately encouraged and supported only local initiatives. From 2014 on, with the same coalition partners and another Christian democrat Minister of Education, the new government made a complete U-turn and launched an extremely ambitious plan called *Pacte pour un enseignement d'excellence* (Pact for excellence in education).

The main goal of the Pact is to "enhance the quality of teaching and education for all students". All stakeholders have been involved in collaborative work on several topics since 2015. The process began in 2015 with two reports: an overview of the current situation, heavily relying on international assessments, and a report on knowledge and skills for the twenty-first century. Almost all topics related to education have been under scrutiny and some major reforms have been adopted, while others either definitely will or may be in the coming years, because the Pact goes beyond the term of the current government (2019). Some reforms are anticipated with long-term goals and gradual implementation.

Among the major topics, one might mention:

- A substantial increase in the resources allocated to kindergarten
- a comprehensive school from grade 1 to grade 9
- The halving of grade repetition by 2030
- An update and revision of all curriculums, at all levels and in all subjects
- A huge bank of validated tools to help teachers adapt their teaching for all levels and all subjects
- In terms of governance, a shift towards a system giving more autonomy to schools but making them accountable.

For people interested in more details, the five strategic axes of the Pact are summarised in a document available (in French only) at http://www.pactedexcellence.be/index.php/lessentiel-du-pacte/. Although the need to strengthen teacher training – initial and in-service – is highlighted in the Pact, a reform of teachers' initial training is not part of the Pact, because this dossier is the responsibility of another Minister – the Minister of Higher Education – controlled by the other coalition party (the socialists).

It is far too soon to anticipate what could be the effects of the Pact. For sure, this is an extremely ambitious multidimensional plan that tackles most of the systemic weaknesses of the education system in FS Belgium at the same time, which is a huge challenge. The time span of implementation is lengthy: some measures have already been adopted, such as the new monitoring of schools, while others are scheduled for the long term (i.e. the comprehensive school, gradually starting in 2021 with a cohort of 5-year-olds).

One group has just started work on how to redefine the system of national assessments (those leading to certificates and the diagnostic ones) in parallel with international assessments and in congruence with the new comprehensive school system. This group has to meet several challenges. In my view, one critical decision is what should be assessed at the end of the comprehensive school (grade 9): should the assessment focus only on minimal competencies defined in the standards, or should it include a broader range of knowledge and skills, and different levels to achieve in order to access different tracks? The direction that is taken will clearly have a backwash effect on the meaning of this new comprehensive school. A second focus of attention is who will be in charge of the development of national assessments in the future. So far they have been developed by teachers supervised by inspectors and with very limited resources. This system has clearly reached its limits and it is time to switch to a more professional approach matching the high ambitions of the Pact in terms of excellence. Excellence in teaching also means excellence in assessment policies and high quality of testing instruments. In this regard, decisive progress must be achieved.

References

Baye, A., Fagnant, A., Hindryckx, G., Lafontaine, D., Matoul, A., & Quittre, V. (2009). Les compétences des jeunes de 15 ans en Communauté française en sciences, en mathématiques et en lecture. Résultats de l'enquête PISA 2006. *Cahiers des Sciences de l'Éducation, 29–30*, 1–245. Retrieved from http://hdl.handle.net/2268/19520.

Demeuse, M., & Monseur, C. (1999). Analyse critique des indicateurs déterminant l'attribution des moyens destinés à la politique de discrimination positive en Communauté française de Belgique. *Mesure et évaluation en éducation, 22*(2–3), 97–127.

Demonty, I., Blondin, C., Matoul, A., Baye, A., & Lafontaine, D. (2013). La culture mathématique à 15 ans. Premiers résultats de PISA 2012. *Cahiers des Sciences de l'Éducation, 34*, 1–28. Retrieved from http://www.aspe.ulg.ac.be/Files/premiers_resultats_pisa_2012__cahiers_34_.pdf.

Dupriez, V., & Maroy, C. (2003). Regulation in school systems: A theoretical analysis of the structural framework of the school system in French-speaking Belgium. *Journal of Education Policy, 18*(4), 375–392.

Fédération Wallonie-Bruxelles. (2016). *Les indicateurs de l'enseignement*. Bruxelles: Service général du pilotage du système éducatif. Edition 2016. Retrieved from http://www.enseignement.be/index.php?page=0&navi=2264B13

Friant, N., Demeuse, M., Aubert, A., & Nicaise, I. (2008). Les politiques d'éducation prioritaire en Belgique: deux modes de régulation des effets d'une logique de marché. In M. Demeuse, D. Frandji, D. Greger, & J.-Y. Rochex (Eds.), *Les politiques d'éducation prioritaire en Europe. Conceptions, mises en œuvre, débats* (pp. 85–133). Institut national de recherches pédagogiques: Lyon.

Lafontaine, D. (1996). *Performances en lecture et contexte éducatif*. Bruxelles: De Boeck.

Lafontaine, D. (2014). A petits pas dans la bonne direction. *Traces de changements, 215*, 4–5. Retrieved from http://hdl.handle.net/2268/173350.

Lafontaine, D., & Blondin, C. (2004). *Regards sur les acquis des élèves en Communauté française Apports des enquêtes de l'I.E.A, de PISA et des évaluations externes*. Bruxelles: De Boeck.

Lafontaine, D., Baye, A., Burton, R., Demonty, I., Matoul, A., & Monseur, C. (2003). Les compétences des jeunes de 15 ans en Communauté française en lecture, en mathématiques et en sciences. *Cahiers du Service de Pédagogie expérimentale, 13–14*, 1–230.

Lafontaine, D., Dupont, V., & Schillings, P. (2017). Pratiques d'enseignement et compétences en lecture des élèves: qu'évaluent les enquêtes internationales et que peuvent en apprendre les enseignants? In M. Bianco & L. Lima (Eds.), *Comment enseigner la compréhension de lecture* (pp. 63–81). Paris: Hatier.

Lafontaine, D., Dupont, V., & Schillings, P. (2018). *Teaching and assessing reading literacy in primary education: Identifying patterns of differences between English, German and French-speaking education systems (PIRLS 2016)* (*IEA Compass: Briefs in Education, 4*). Amsterdam: IEA.

Lafontaine, D., Bricteux, S., Hindryckx, G., Matoul, A., & Quittre, V. (2019). *Performances des jeunes de 15 ans en lecture, mathématiques et sciences. Premiers résultats de PISA 2018 en Fédération Wallonie-Bruxelles*. Université de Liège: Analyse des systèmes et des pratiques d'enseignement.

Mons, N. (2007). *Les nouvelles politiques éducatives*. Paris: PUF.

Monseur, C. (1997). *L'enseignement des sciences est-il dans le 36e dessous?* Unpublished document. Université de Liège, Liège.

Monseur, C., & Lafontaine, D. (2009). L'organisation des systèmes éducatifs: quel impact sur l'efficacité et l'équité? In V. Dupriez & X. Dumay (Eds.), *L'efficacité en éducation, promesses et zones d'ombre* (pp. 185–219). Bruxelles: De Boeck.

Monseur, C., & Lafontaine, D. (2012). Structure des systèmes éducatifs et équité: un éclairage international. In M. Crahay (Ed.), *Pour une école juste et efficace* (pp. 145–173). Bruxelles, Belgique: De Boeck.

OECD. (2011). *PISA 2009 results. Overcoming social background. Equity in learning opportunities and outcomes*. Paris: OECD.

OECD. (2016). *PISA 2015 results. Excellence and equity in education*. Paris: OECD.

Quittre, V., Crépin, F., & Lafontaine, D. (2018). Les compétences des jeunes en sciences, en mathématiques et en lecture. Résultats de PISA 2015 en Fédération Wallonie-Bruxelles. *Cahiers des Sciences de l'éducation, 37*, 1–188. Retrieved from http://hdl.handle.net/2268/218401.

Chapter 8
International and National Assessments in Croatia

Michelle Braš Roth

This chapter refers to the monitoring and assessment of pupils' achievements in the Republic of Croatia, starting with a brief chronological review and describing some of the changes that have taken place in the education system over the last 25 years. For a decade, Croatia has been introducing new ways of external evaluation in education at the national level and has participated in several international studies to compare learning outcomes and the quality of the education system with other countries. During the first two decades of the twenty-first century, the quality of education and the constant monitoring of educational achievements has become imperative not only for teachers but also for employers and society as a whole. Teachers will necessarily have to change the teaching methods and have permanent training for new challenges in their profession, especially to meet the constant changes and needs of the labor market, to better prepare students for their first occupation or for continuing their education and taking their role of active citizens in a modern democratic society. Consequently, monitoring students' assessment and achievements becomes a generator of key changes in education systems and determines significant trends in the development of educational practice in the future.

Introduction to International and National Assessment Context and Its History in Croatia

Since the dissolution of the Yugoslav Federation during the 1990s, the Republic of Croatia, like the rest of the former socialist countries, has started the process of transition from communism to pluralist democracy, the free market, and the unification with the European Union. This also initiated changes in the educational system.

M. Braš Roth (✉)
National Centre for External Evaluation of Education, Zagreb, Croatia
e-mail: mbc_studio@yahoo.com

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_8

Unfortunately, at the beginning of this process, the country went through a defensive war that destroyed and damaged many industrial buildings, cultural, religious, and educational institutions and slowed the implementation of all planned reforms. The newly established government had to take care of hundreds of thousands of displaced persons and later refugees from Bosnia and Herzegovina, rebuild what was destroyed, create new democratic institutions and democratize society as a whole, privatize public and material assets, restructure the economy, and work on joining Euro-Atlantic integrations.

The Republic of Croatia entered the twenty-first century with an education system that had to be reformed, bearing in mind the Croatian and European context. It was necessary to plan the development of education according to European standards and design a system similar to those in developed European countries. This was supposed to contribute to the goal of shared living and sharing responsibilities in the united Europe and strive to take part in the creation of new European school models in the future.

Over the past two decades, much attention has been directed to the educational reform, but unfortunately, important and desired changes in structure and quality have not been achieved. There were three attempts to introduce a new concept between 1990 and mid-1999, when extensive reform was proposed. In 1991, a public tender for changes in the education system was opened; in 1993, the document "New Croatian School: General Concept of Education in the Republic of Croatia" was proposed, and in 1995 the Croatian Education Development Program was proposed.

From 1990 to 2000, there were various partial changes in the structure of the educational system and in the curriculum. The primary school curriculum was reformed to comply with the demands of a free and democratic society, rejecting ideology and adhering the principles of pluralism and modernization. About 180 vocational-educational centers were transformed into 399 secondary schools (grammar schools and 3- or 4-year vocational schools) with new curricula and syllabuses drawn up for all types of secondary schools and occupations. Furthermore, the dual system was introduced in crafts training, religious instruction became an elective subject, and information technology began to be introduced in schools, both in primary and secondary education.

In the period from 1990 to 2000, several laws were adopted to regulate the legal, financial, and organizational aspects of each subsystem (preschool, elementary, and secondary education). These legal changes have enabled educational pluralism such as opening of private and international schools and the introduction of alternative programs such as the Waldorf and Montessori schools. Furthermore, local communities started to cofinance primary and secondary schools, and local self-government bodies and parents gained a greater participation in the decision-making process. It also enabled an adequate education for national minorities as well as greater autonomy for schools and teachers.

In 2000, the Ministry of Education and Sports started a public debate on the Fundamentals of the Educational System in the Republic of Croatia, a document reflecting the government's educational reform program and placing education at

the focus of interest of the broadest professionals and wider public in Croatia. Croatia has opted for an education system similar to those in economically, scientifically, and technologically developed countries, but before reforms could be planned and implemented, Croatian education had to be analyzed from different aspects and compared with education in other countries.

Therefore, the National Center for External Evaluation of Education was established, which has immediately started with a series of activities in the field of measurement of educational outcomes at the national level. In addition, Croatia joined several international studies with the aim of monitoring pupils' achievements and the quality of the education system led by International Association for the Evaluation of Educational Achievement (IEA) and Organization for Economic Co-operation and Development (OECD):

- Programme for International Student Assessment (PISA 2006, PISA 2009, PISA 2012, PISA 2015, PISA 2018)
- European Survey on Language Competencies – ESLC (SURVEYLANG 2011)
- Progress in International Reading Literacy Study (PIRLS 2011)
- Trends in International Mathematics and Science Study (TIMSS 2011, TIMSS 2015, eTIMSS 2019)
- International Computer and Information Literacy Study (ICILS 2013)
- Teaching and Learning International Survey (TALIS 2013, TALIS 2018)
- International Civic and Citizenship Study (ICCS 2016)

Although the results did not receive much media coverage, many people were very much surprised with the outcomes. Participating schools received feedback on the success of their students and the placement of Croatia at international scales and started to discuss them. Within the PISA study, thousands of parents of sampled students received a personalized report on their child's competences in relation to the other 15-year-olds in the country, but also around the world, the description of levels of competences that could be compared to their school grades. This kind of feedback to schools and parents ensured not only a very high response rate of the participants for next PISA cycles but also increased the motivation for teachers to learn more about teaching methods in other countries in order to improve their teaching and assessment methods as well. The results of these surveys have deepened the interest of the society for quality education and faster changes in the system of compulsory and secondary education.

The Croatian Parliament adopted the *Strategy of Education, Science and Technology* in 2014, in which the focus of the planned changes was set on the basis of accurate external evaluation indicators, both nationally and internationally. Based on this strategy, preparations for a complete curricular reform have begun, and this education reform is undoubtedly one of the most demanding projects that Croatia faces at the beginning of the twenty-first century.

The current Croatian education system begins in preschool institutions, which cover children from the age of 6 months to their start of schooling. They can be run by local authorities or private nursery schools (legal persons, religious communities, and others). Elementary schools can also provide shorter preschool programs.

Elementary education, which lasts 8 years, is compulsory for children who are six and a half or over. There is an adult elementary education system for those who do not complete primary education by the age of 15. Secondary education is optional and is divided according to curricula into gymnasiums, vocational schools (technical, industrial, and craft based), or art schools (music, dance, art). Gymnasiums provide a comprehensive syllabus which lasts 4 years and includes a final examination, the state matura. Programs in vocational and art schools last from 1 to 5 years, and usually end with the production of a final assignment, but it is also possible to sit the state matura if pupils have completed 4 years of secondary education. Since 2010, state matura results have been the basis for entry to higher education institutions. Elementary and secondary education in state schools is free. Higher education institutions are divided into polytechnics, colleges of applied science, faculties, and art academies. All study programs were aligned by 2005 with the requirements of the Bologna Process as part of the creation of a European system of higher education.

International and National Assessments in Croatia Today

The system of external evaluation in education is one of the strategic objectives of Croatian education and is described in detail in the document "*Education Development Plan 2005–2010*" issued by the Ministry of Science, Education and Sports (2005). As a mechanism for the objective monitoring of the education system, it is one of the key factors which influences the improvement of the quality of the education system as a whole.

According to the education law, schools are required to participate in the external evaluation and use the results of the self-analysis and self-assessment as part of the external evaluation for the purpose of continuously improving the quality of education and their work.

External evaluation of education in Croatia is based on standardized examinations in major subjects such as mathematics and science conducted by the National Center for External Evaluation of Education (NCEEE) since 2006.

In cooperation with the Institute of Social Sciences Ivo Pilar during the school year 2006/2007, National Center for External Evaluation of Education organized and conducted the first experimental external evaluation of the educational achievements of elementary school students. The next school year an external evaluation of the educational achievements of all fourth and eighth grade students in elementary schools was carried out, where the entire population of 46,556 students was included. The elementary schools thus entered the system of external evaluation of the students' educational achievements. The basic objective of this project of external evaluation of the educational achievements of the fourth and eighth grades of elementary schools was to determine the level of acquired knowledge, skills, and abilities of students in particular subjects and curricula and to examine the pupil's abilities to interdisciplinary link contents from different subjects.

The National Center for External Evaluation of Education continued to develop a system for external evaluation of educational outcomes in primary schools and developed strategies, i.e., multiyear development projects such as Development and Strategy of National Examinations (2008–2009) and Development of final exams at the end of educational cycles (2011–2015). As a result of work on improvement of national exams, two more assessments were carried out in the following years: external evaluation of Biology in eighth grades (school year 2010/2011) and National Examinations in Mathematics in eighth grades (2011–2014).

Currently, Croatia is continuing to develop national exams which are standardized and designed to determine achievements regarding basic knowledge and skills of students in the most important subjects within key parts of educational cycles or/and at the end of compulsory education. National exams are based on the in-depth analysis of previously conducted tests, the psychometric characteristics of assessment tasks, and the obtained indicators. These examinations should provide a reliable insight into the functioning of the education system in order to determine the necessary measures of quality improvement of the system on the basis of these results.

Preparations for the introduction of graduation state exam as the final exam at the end of secondary education started with national secondary school examinations as well: national exams in first grade of gymnasium school programs in 2006, in second grades in 2007, in first grade of 4-year vocational programs in 2007, and in third grades of grammar school and 4-year vocational programs in 2008.

In 2008, the Croatian Parliament adopted the Law on Primary and Secondary Education which established the legal basis for the introduction of graduation state exams (matura) into the Croatian secondary school system. The Ministry of Science, Education and Sports prepared a rulebook on Graduation State Exams, which sets out the detailed rules for examining. After a comprehensive education campaign conducted in all 4-year secondary schools, the first graduation state exams were carried out in 2009/2010 school year.

All students, completing their education in gymnasiums, are obliged to take the graduating state exams to complete their secondary education. It is compulsory for students of 4-year vocational high schools, who successfully complete the fourth grade and want to continue their education at one of the universities in the Republic of Croatia as well.

By introducing graduation state exams, national standards for evaluating school achievements of students at the end of 4 years of secondary education were defined, and greater objectivity in evaluating students' achievements was ensured. The students' results at graduation state exams show objective student knowledge and achievements, and individual student exam results are also a form of evaluation for enrollment in higher education institutions. However, there is still a lot of issues regarding subjectively or qualitatively assessed competencies which need to be solved.

In 2017, the National Center for External Evaluation of Education began with the implementation of the project Development of the National Examination System, which is based on two major factors: the development of a national examination

system and the development of an item bank with metric and content characteristics of the items that will be available to teachers and students. Therefore, during 2018, in the pilot phase of this project, a software for building an item bank and test design, online test delivery, and marking was purchased. In April 2018, the first online assessment for primary school students in the field of Information Technology and Physics was successfully conducted. Most probably, this project will improve the teachers' role in the process of assessment in the near future since more and more teachers are participating in the development of the testing items, analyzing and interpreting of the results, and many of them are interested to attend seminars organized by the National Center for External Evaluation of Education. Croatia has continued with the implementation of international research and is currently participating in PISA, PIRLS, TIMSS, and ICCS.

Programme for International Student Assessment (PISA)

The PISA 2015 cycle is the sixth cycle of the Organization for Economic Cooperation and Development's survey and the fourth one in which the Republic of Croatia participated. It was the second time that science literacy was examined as the main domain, while reading literacy and mathematical literacy were examined as secondary domains. Additionally, it also examined the students' competencies in collaborative problem solving as an innovative domain.

Seventy-two countries participated in the study and a total of 540,000 students were tested, representing about 29 million 15-year-old students in participating countries. The Republic of Croatia was represented by 5809 15-year-old students from 158 secondary and 2 primary schools. Testing of students was performed for the first time exclusively on the laptops in all test domains. In addition to the cognitive test, students, their parents, and school principals filled out various questionnaires providing a large amount of background data linked to educational achievements.

Croatian Results

The average score of Croatian students in **science literacy** (Braš Roth et al. 2016) is 475 points, which is below the OECD average and is ranked 37th in the international rankings. By comparing Croatia's average results with the results of other countries, the achievements of Croatian students do not differ significantly from the achievements of students from neighboring Italy or Hungary, as well as Lithuania, Argentina, and Iceland. Compared with the results of PISA 2006 (Braš Roth et al. 2008), when science literacy was also the main test domain, there was a significant fall in the average achievement of Croatian students. On average, every 3 years, the achievements of Croatian students drop by about 5 points.

When it comes to knowledge and abilities on the scale of science literacy, almost a quarter (24.7%) of 15-year-old Croatian students did not reach proficiency Level 2, meaning they do not possess the basic science competencies necessary for everyday life. At the highest proficiency levels (levels 5 and 6), there were only 4% of Croatian students. When the distribution of Croatian students on a scale of science literacy in this cycle is compared with the previous three cycles, it can be noticed that second level in 2006 was not reached by 17% of students, 18.5% in 2009, 17.2% in 2012, and in this cycle by 19.2% of students. At the fifth and sixth levels in 2006, there were 5.1% students, 3.7% students in 2009, 4.6% students in 2012, and 4% of students in 2015. According to an analysis of the gender differences in Croatia, as well as in OECD countries, no statistically significant differences were found between girls and boys with respect to science literacy.

In **reading literacy** (Braš Roth et al. 2010) Croatian students achieved a below-average score of 487 points and were ranked 31st. By comparing the average results in reading literacy with the results in PISA 2009, Croatia has shown a trend of improving average results. In the 6-year period, Croatia has increased the average score by 11 points. Regarding the comparison of the distribution of Croatian students by the level of reading literacy in the 2015 cycle with the previous three cycles, it can be noticed that percentage of students below second level is slightly declining (21.5% in 2006, 22.4% in 2009, 18.6% in 2012, 19.9% in 2015) and increasing at the fifth and sixth levels (in 2006 there were 3.7% students, 3.2% students in 2009, 4.4% in 2012, and 5.9% in this cycle). When it comes to gender differences in the OECD countries, in this PISA cycle, girls have scored in average 27 points better than boys. In Croatia, this difference in favor of girls is 26 points and has been significantly decreasing compared to previous cycles: in PISA 2006, girls were better than boys by 50 points, in the PISA 2009 cycle by 51 points, and in the 2012 PISA cycle by 48 points. It is also important to notice the representation of boys and girls at the lowest and highest levels of reading literacy. In the group of students who did not reach Level 2, there are 25% boys and 15.1% girls, while at the highest levels of knowledge and skills, there are 4.7% boys and 7.0% girls.

Mathematical literacy (Braš Roth et al. 2013) was the weakest domain for Croatian students with an average score of 464 points which is below the OECD average. In this domain Croatia was ranked 41th. Compared to 2012, when mathematical literacy was the main test area, Croatian students achieved a weaker score by 7 points, but this difference is not statistically significant. Moreover, changes in the achievements of Croatian students in mathematical literacy since 2006 did not prove to be significant. However, it is important to notice that almost one third of the Croatian students could not reach the second level in mathematical literacy, which is the basic level of mathematical competencies. It can be noticed that second level in 2006 was not reached by 28.6% of Croatian students, 33.2% in 2009, 29.9% in 2012 and 29.9% in 2015. At the fifth and sixth levels in 2006 there were 4.8% of students, 4.9% of students in 2009, 7.0% of students in 2012, and 5.6% of students in the 2015 cycle. As far as gender is concerned in OECD countries, boys are on average more successful in math than girls by 8 points. In Croatia, boys also have significantly better results in math than girls, and this difference is 13 points. If this

difference is compared to the previous PISA cycles, the difference in boys' benefit has not changed significantly (in 2006 it was 13 points, in 2009 it was 11 points and 12 points in 2012). Given the presence of boys and girls at the lowest and highest level of mathematical literacy, it is noted that the group of students who did not reach Level 2 comprised 30% boys and 33.9% girls, while at the highest level of knowledge and abilities 7.1% were boys and 4.1% were girls.

In the PISA 2015 cycle, an innovative domain was developed to measure students' competences in **collaborative problem solving** (Braš Roth et al. 2014b). Evaluation of the ability to collaborate to solve the problem was carried out in 52 participating countries. In the overall ranking of 52 countries that assess the competencies in collaborative problem solving, Croatia is ranked 32nd. The average Croatian student score was 473, which put Croatia in a group of countries with a statistically lower score than the OECD average. The comparison between countries has shown that the results of Croatian students are not statistically significant compared to the results of Italy, Russia, Hungary, Israel, and Lithuania.

In order to better interpret students' achievements in evaluating collaborative problem solving, the overall scale is divided into five levels of knowledge and achievement. Below the first level, there are 6.6% of Croatian students, while more than a quarter of Croatian students (28.7%) are at Level 1 and only have basic skills for collaborative problem solving. The highest percentage of students (41.8%) met Level 2, which means they were able to contribute to a collaborative problem solving and to negotiate with other members of the team on procedures how to solve the problem. At the third level there were 20%, while at the highest level there only 2.4% of Croatian students, which was ten times less than in first ranked Singapore. In the PISA 2012 cycle, which focused on mathematical literacy as the main test domain, financial literacy was tested for the first time as a secondary domain of assessment.

Financial literacy (Braš Roth et al. 2014a) was tested in 18 of the 65 participating countries, including the Republic of Croatia. On the overall scale of financial literacy, Croatia was ranked 14th. The average Croatian score was 480 points, which placed Croatia in a group of countries with a significantly below the OECD average. Cross-country comparison showed that the score of Croatian students did not differ statistically from the results of the United States, the Russian Federation, France, Slovenia, Spain, Israel, and Slovakia. Students' achievements are presented on the overall financial literacy scale, which includes five levels of knowledge and abilities. On this scale, 15.5% of Croatian students did not reach Level 2 or have basic skills in financial literacy. On the other hand, slightly more than 10% of Croatian students achieved the highest level of excellence level (fifth level) in this area.

By comparing results by gender, in almost all countries, including Croatia, there was no difference between boys and girls. Correlations with achievement in reading literacy and mathematical literacy on the one hand and with achievement in financial literacy on the other hand are extremely high and positive. In other words, students that achieve better results in mathematical and reading literacy also achieve better results in financial literacy. Greater correlation was found with mathematical literacy than reading literacy.

How to Go Further?

To continuously improve and enhance the quality of the education system, it is extremely important to develop a model which uses the results of periodic external examinations to monitor the achievement of educational outcomes at the national level, as well as the model for the use of indicators gained under international educational studies. The educational authorities in the Republic of Croatia devoted a lot of attention to the continuous upgrading of quality and excellence in education. This is determined by many factors, including digitization of elementary and secondary education, raising teachers' level of expertise, self-evaluation of schools, and national examinations as an external evaluation. Much has been done especially in the field of external evaluation of education and self-evaluation of schools in the last decade. Unfortunately, comparative results and other indicators gained in international research have not been used enough to plan or improve educational work so far. Perhaps the best indicator of the inadequate use of the obtained results is precisely the fact that, for example, the PISA results in all three test domains have hardly changed in three consecutive PISA cycles. Croatia is not the only country where there is no significant change in the average results of any test domain since the country started participation in the survey. The average results in the test domain have almost remained the same for 10 years.

The reason for this can be found in the fact that no significant changes in educational background occurred during that period, which could lead to substantial changes in educational outcomes. It is also possible to look for the reason in the fact that education policy did not sufficiently clarify goals in relation to indicators of international research. Although teachers and educational professionals are well informed about the relatively low average score of Croatian students, there is a lack of a defined action plan to improve such results in the following cycles of various international studies in which Croatia participates in order to monitor the quality of the education system. However, as in other education systems, there is certainly a certain percentage of teachers in Croatia who are continually trying to improve their work and student achievements, regardless of politically defined educational decisions and drafted reforms. They are well informed about contemporary educational practices and ways of monitoring achievements, many of them with the mentor or counselor status actively cooperate with the Teacher Training Agency and the National Center for External Evaluation of Education and through seminars and different working groups also contribute to the professional development of their colleagues.

In addition, insufficient attention has been given to secondary analysis and linkage with national, as well as other international research. For example, it is interesting to note that Croatian students at the age of 10 as participants in the PIRLS 2011 study achieved above-average scores on the international scale, but at the age of 15 in the PISA survey, they turn out to be below average three times in a row. Croatian educational experts have not yet given an answer to the question of what may have happened during the last 4-year period of the compulsory education

where students' competence in reading literacy and reading engagement suddenly falls and what reactions or curricular changes are needed to improve the results.

National exams are more oriented in measuring educational outcomes according to the national curriculum, but the main science or mathematical competencies that are assessed in international research such as PISA and TIMSS should be very similar to most prescribed educational content in different countries, only the questions are placed in the actual life context with the aim of assessing students' ability to apply this knowledge in everyday life situations. It is therefore important that the development of national exams methodologically approaches this international criteria, especially in the development of test instruments.

Learning for a test or preparing students for key national tests poses a risk of over-memorizing facts instead of adopting basic concepts and developing skills in applying the acquired knowledge in everyday life situations. PISA is a kind of assessment where the competences that are measured by the framework and the exam questions themselves are clearly defined and the level of knowledge and skills that a learner needs to possess at a certain level as well. According to that, teachers' better knowledge of the PISA concept could also serve for planning adequate teaching methods due to efficient development of the determined competencies. In that case, the process of monitoring of the students' progress is used directly and efficiently for constant modification of teaching methods and metacognitive data for advancing the learning process itself.

Worldwide, recent lists of twenty-first-century skills are extended to more and more soft skills, behavioral skills, contextual learning skills, creative and critical thinking, self-management, cultural and health awareness, civic and entrepreneurial literacy, etc. The demand for developing these skills comes from the global labor market and requires changes parallel to changes in the global economy and from conditions in modern, multicultural societies. Educational systems should respond as quickly and efficiently to such requirements. That is why changes in national curricula occur more and more often, and these changes should be based on precisely measured and analyzed data. Therefore, evidence-based policy expects increasingly more from the monitoring of student achievement, so even this process of evaluation is rapidly and ever changing.

Croatian Strategy of Education, Science and Technology defined also the development and establishment of an ICT system for digital learning outcomes as one of the measures for educational system improvement. Computer-based assessment was already successfully used in ICILS and PISA survey a few years ago. Use of technology in monitoring of students achievement is not only to measure ICT literacy skills, but should also be extended to national exams and everyday use in the learning process. This type of assessment opens the possibility for students' knowledge and skills to be tested through interactive tasks, using simulations and other innovative item types or using different sources of information during the test session. Collaborative problem solving as a competence was already tested in the PISA 2015 cycle; critical thinking and creativity are just some of the competencies that are already developing for the next cycles.

Computer-adaptive testing allows the teacher and the student to monitor their level of performance and thus obtains faster and better feedback about the parts of the lesson to be repeated.

Web-based testing is also one of the solutions for very convenient test management, and Croatia has already started to pilot that option of monitoring achievement. However, there will always be competencies and skills that only teachers with their professional knowledge can and must evaluate.

References

Braš Roth, M., Gregurović, M., Markočić Dekanić, A., & Markuš, M. (2008). *PISA 2006 – Science competencies for life*. Zagreb: Nacionalni centar za vanjsko vrednovanje obrazovanja – PISA centar. https://mk0pisancvvooeeubi4r.kinstacdn.com/wp-content/uploads/2018/09/IZVJESTAJ-CJELOVITI_PISA2006.pdf.

Braš Roth, M., Markočić Dekanić, A., Markuš, M. i., & Gregurović, M. (2010). *PISA 2009. Reading competencies for life*. Zagreb: Nacionalni centar za vanjsko vrednovanje obrazovanja. https://mk0pisancvvooeeubi4r.kinstacdn.com/wp-content/uploads/2018/05/IZVJESTAJ_PISA2009_press.pdf.

Braš Roth, M., Markočić Dekanić, A., Markuš Sandrić, M. i., & Gregurović, M. (2013). *PISA 2012: Mathematical competencies for life*. Zagreb: Nacionalni centar za vanjsko vrednovanje obrazovanja. https://mk0pisancvvooeeubi4r.kinstacdn.com/wp-content/uploads/2018/05/IZVJESTAJ_PISA2012_matematicke_46_finn.pdf.

Braš Roth, M., Markočić Dekanić, A., Gregurović, M., & Ružić, D. (2014a). *PISA 2012: Financial literacy*. Zagreb: Nacionalni centar za vanjsko vrednovanje obrazovanja. https://pisa.ncvvo.hr/wp-content/uploads/2018/05/IZVJESTAJ_PISA2012_Financijska_26_finn_2.pdf.

Braš Roth, M., Markočić Dekanić, A., & i Gregurović, M. (2014b). *PISA 2012: Problem solving competencies*. Zagreb: Nacionalni centar za vanjsko vrednovanje obrazovanja. https://mk0pisancvvooeeubi4r.kinstacdn.com/wp-content/uploads/2018/05/IZVJESTAJ_PISA2012_problem-solving_15_zaTisak.pdf.

Braš Roth, M., Markočić Dekanić, A., & Markuš, M. (2016). *PISA 2015 – Science competencies for life*. Zagreb: Nacionalni centar za vanjsko vrednovanje obrazovanja. https://pisa.ncvvo.hr/wp-content/uploads/2018/05/PISA-2015-kb.pdf.

Law on Primary Education. Official Gazette No. 59/90, 26/93, 27/93 and 7/96. Zagreb: Narodne novine.

Law on Secondary Education. Official Gazette No. 19/92, 27/93 and 7/96. Zagreb: Narodne novine.

Nacionalni okvirni kurikulum za predškolski odgoj i obrazovanje te opće obvezno i srednjoškolsko obrazovanje. (2010). Zagreb: Ministry of Science, Education and Sports.

Report on external evaluation in elementary schools. (2009). Zagreb: Nacionalni centar za vanjsko vrednovanje obrazovanja.

Strategy on Science, Education and Technology. Official Gazette No. 124/2014. Zagreb: Narodne novine.

Chapter 9
The Evolution of National and International Assessment in England

Liz Twist

Education in England

Within the United Kingdom, education policy and delivery is devolved to the governments of England, Northern Ireland, Scotland and Wales. In England, responsibility for the administration of education and the setting of educational standards and regulations is held by the education ministry, known as the Department for Education. All students are entitled to free education up to age 18. Compulsory schooling is from age 5, but most children experience preschool education, some of which is funded by the state. Whilst compulsory schooling ends at age 16, students must continue in some form of education, employment or training until age 18.

Early Development of National Assessments in England

National assessments have evolved in England since the introduction of the first national curriculum in 1988. Prior to this, there was a system of national monitoring surveys on a sample basis ('Assessment of Performance Unit' surveys). Individuals took assessments at age 16, marking the end of statutory schooling; at age 18, for a minority of students, there were examinations which determined access to higher education. In a few localities, there was a test at age 11, providing successful students with access to the highly academic education provided by grammar schools at the secondary level (ISCED levels 2 and 3).

Whetton (2009) outlines the political influences that led to the education reform and the first national curriculum in 1988. In essence, despite increases in funding leading to, for example, lower pupil-teacher ratios, there was thought to be limited

L. Twist (✉)
National Foundation for Educational Research, Slough, UK
e-mail: l.twist@nfer.ac.uk

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_9

evidence of higher standards being reached. Alongside the development of the national curriculum, moves were made to devise a system of national assessment. Simultaneous reform of these two building blocks of the education system – a national curriculum and a national assessment model – was not to be undertaken again until 2014.

The 1988 national curriculum identified three 'core subjects' – English (comprising speaking and listening, reading and writing), mathematics and science. These were the subjects that became the focus of the emerging assessment system, the blueprint for which was developed by the Task Group on Assessment and Testing, known as TGAT (DES and WO 1988). From this time, the curriculum began and continues to be organised in phases, known as 'key stages' with the main point of transition (from primary to secondary education, ISCED levels 1 and 2/3) at age 11:

Key stage 1, years 1–2, age 5–7.
Key stage 2, years 3–6, age 7–11.
Key stage 3, years 7–9, age 11–14.
Key stage 4, years 10–11, age 14–16.[1]

In 1988, new public examinations were taken at the end of key stage 4. From this date, statutory assessments were introduced at the end of the other three key stages in a gradual manner, starting with arguably the most challenging – tests for 7-year-olds. The overarching purpose of these assessments has evolved from one focusing on the measurement of individual student achievement to one of evaluating the effectiveness of each school; in its current form, implications for students are, in fact, limited. This accountability system, as it is known, is discussed briefly below.

Changes to the National Assessment System in England

National curriculum assessment has evolved since the 1990s.[2] Whilst there has always been a requirement to produce scores or grades for individual students, there is now a focus on valuing both attainment and progress equally by identifying students' starting points and comparing the amount of progress made by students from the same starting point. The national assessments at age 11 are considered to be 'high stakes' for schools – they are one element considered by the schools' inspectorate Ofsted as a measure of school effectiveness – but for the most part they have no consequences for individual students.

Following a public consultation, the government's primary assessment policy position was published in 2017 (DfE 2017b). From 2020, the intention is to introduce a new assessment of children on entry to school (the 'reception baseline

[1] This particular structure continues to the present time.

[2] For a summary of the changes up to 2008, see Whetton (2009).

assessment') and use this as the basis for measuring the effectiveness of primary schools. At present, the plan is that in 2027, primary school performance will be based on students' attainment in reading and mathematics tests at age 11 and on the amount of progress made, compared to students' with the same starting point in the reception baseline assessment, over the 7 years of primary education. The introduction of an assessment of children on entry is controversial. Some are opposed in principle to the assessment of 4-year-olds; others welcome the opportunity to recognise the progress made throughout primary school including the vital early years.

This is the most fundamental change to the assessment system since the 1990s but there have been other significant changes. The status of summative teacher assessment[3] has gradually been diminishing. Once the reception baseline is established, there will be no statutory assessment at age 7, and from 2018, there is no longer any requirement to undertake a teacher assessment of students' reading and mathematics at age 11 (where there are statutory tests in these subjects). A completely new assessment was introduced in 2012 – an assessment of early reading knowledge – and this is discussed in the next section.

Phonics Screening Check

Of particular relevance to the evolution of reading assessment in England is the introduction of the 'Phonics Screening Check'. This new statutory assessment was introduced in 2012 with the aim of ensuring that a phonics-based approach to the teaching of reading was adopted in the early years of schooling in England. Pupils in year 1 (aged 5–6) are assessed at the end of the school year, and the test framework (DfE 2012) states that the purpose of the assessment is 'to confirm that all children have learned phonic decoding to an age-appropriate standard'. Students are asked to decode words which become, in phonic terms, increasingly complex. At the start, students are presented with 'pseudo-words', i.e. words invented for the purpose of the check that are phonically regular and plausible but do not exist in English. Examples in 2018 included 'reb' and 'zook'. In the next part, students are asked to read English words – examples from 2018 are 'dart' and 'gift'. Finally students are asked to decode words containing more complex structures (such as 'splote' and 'modern' in 2018).

There is a threshold on the check (since 2012 this has always been 32 as the tests are trialled and constructed to be equivalent each year). Pupils who score below this (or do not take the check) are expected to take the following year's check, after additional support. In 2012, 58% of pupils achieved at least 32 marks and so met the expected standard. This proportion has risen each year since and in 2018 the figure was 82% (DfE 2018a).

[3] Teacher assessment is a criterion-based judgment made by teachers using a broad range of evidence from across the curriculum and knowledge of how a pupil has performed over time and in a variety of contexts. It is carried out as part of teaching and learning.

The final evaluation of the introduction of the phonics screening check (PSC) (Walker et al. 2015) suggested that it had led to an increase in the pace of phonics teaching and to more systematic teaching of phonics. However, it was not possible to attribute any improvement in students' literacy attainment to the check due to methodological issues (there was no control group and a range of other literacy initiatives were introduced at the same time). There remains some opposition to the check. The United Kingdom Literacy Association submitted evidence to the 2017 enquiry into primary assessment and argued that 'the Phonics Check has contributed very little of value to the reading assessment processes' (UKLA 2017, p. 7). Others argue that what was originally introduced as a 'light touch' diagnostic assessment has become 'a high stakes assessment with schools expected to raise their percentage pass year on year' (Clark and Glazzard 2018, p. 3).

In Australia, researchers such as Buckingham (2016) have encouraged consideration of the PSC, and interest has been shown in this initiative by several states. A trial involving 50 schools in South Australia was undertaken in 2017. Following an independent evaluation (Hordacre et al. 2017), the trial was considered successful and the screening check is now in use across the state. The Federal Education Minister is encouraging wider use in Australia although the rather polarised reaction to his proposition is similar to that in England when the check was first introduced.

Phonics Screening Check and Relationship to PIRLS 2016

When the 2016 PIRLS results were published, the Minister for School Standards in England linked the improvement in the country's performance in PIRLS between 2011 and 2016 to the introduction of the PSC. The significance is that students sampled in PIRLS 2016 were among those who took the first check in 2012 and that the screening check had been initiated and strongly supported by the Minister.

> *The pupils who took part in the [PIRLS 2016] international survey were the first cohort to have taken the Phonics Screening Check in 2012; the cohort to have been taught to read after we changed the law requiring schools to use phonics. The details of these findings are particularly interesting; I hope they ring in the ears of opponents of phonics whose alternative proposals would do so much to damage reading instruction in this country and around the world.* (Gibb 2017)

This perhaps rather overstates the opposition to phonics as a means of teaching reading – opponents tend to focus on the emphasis on phonics to the exclusion of other approaches rather than oppose phonics per se. Castles et al. (2018) suggest that much of this opposition is due to the fact that there has been little attention paid to the discussion of reading development processes *beyond* phonics. The Minister's suggestion of the causal link between improvement on PIRLS 2016 and the introduction of the check was refuted by some (e.g. Clark and Glazzard 2018). England's national report (McGrane et al. 2017, p. 65) details a moderate and statistically significant correlation between score on the phonics screening check and scale score on PIRLS (0.52) – an association but not necessarily a causal relationship. Of all the

assessments introduced on a statutory basis in England, the Phonics Screening Check is the one which was most explicitly designed to influence day-to-day teaching practices – albeit subsequently supported by the content of the 2014 national curriculum. The next section looks at participation in the international large-scale surveys of achievement before a focus on PIRLS, another reading assessment.

England's Involvement in International Assessments

England has a long history of participation in international surveys of achievement, and this sits alongside a statutory national assessment system that has evolved over the past 30 years. England was one of the participating countries in IEA's first international survey ('Pilot Twelve-Country Study'[4]). The success of this survey led to the First International Mathematics Study[5] (FIMS) in 1964, the predecessor of TIMSS, and in which England participated.

In the succeeding years, England has participated in PISA,[6] PIRLS and TIMSS on each occasion the surveys have been administered internationally. England's performance on the three PISA domains (reading, mathematical and scientific literacy) has remained relatively stable with no significant change in the four surveys between 2006 and 2015 (NFER 2016a). In TIMSS, performance in mathematics at grade 4 has not changed significantly since 2007, following significant improvement after the first survey in 1995 (NFER 2016b). There was an improvement between 2003 and 2007 at grade 8. With regard to science, the trend is one of no significant change with one notable exception – the 2011 survey results which saw a significant fall in science attainment at both grade 4 and grade 8. In 2010, statutory science testing across the full year group was abolished at age 11, following the abolition of all statutory testing at age 14 after 2008. This left only mandatory teacher assessment of science. The dip seen in TIMSS 2011 was widely interpreted as a consequence of the reduced status of science in the implemented curriculum due to its diminished role in the national assessment system. Of the three ILSAs, PIRLS is the newest survey and has been administered every 5 years from 2001. Following a dip between the first and second surveys (from an average score of 553 to one of 539), England's performance significantly increased in 2011 (average score of 552) and again in 2016 (559).

Looking ahead, the new curriculum introduced in 2014 is expected to have the effect of raising school standards. The impact of this more demanding curriculum will only be known with the results of the PISA 2021 cycle as that is the first survey that will include students who have experienced a considerable part of their

[4] https://www.iea.nl/pilot-twelve-country-study

[5] https://www.iea.nl/fims

[6] In PISA, the United Kingdom is included in the international rankings and there are subsequently national reports produced for the participating constituent nations (England, Northern Ireland, Scotland and Wales).

schooling with that curriculum. TIMSS 2019 will similarly provide a useful measure of the impact of the increased expectations, especially at Grade 4 as those students will have known nothing other than the 2014 curriculum. The next section will focus on how reading is conceptualised in England's recently reformed national curriculum; how it is defined in PIRLS; how reading is measured in England and in PIRLS and the results reported.

Reading as Defined in England's National Curriculum and in PIRLS

There is no explicit definition of reading in the English national curriculum; rather, the significance of the study of English to life in general and to education in particular is detailed:

> *English has a pre-eminent place in education and in society. A high-quality education in English will teach pupils to speak and write fluently so that they can communicate their ideas and emotions to others and through their reading and listening, others can communicate with them. Through reading in particular, pupils have a chance to develop culturally, emotionally, intellectually, socially and spiritually. Literature, especially, plays a key role in such development. Reading also enables pupils both to acquire knowledge and to build on what they already know. All the skills of language are essential to participating fully as a member of society; pupils, therefore, who do not learn to speak, read and write fluently and confidently are effectively disenfranchised.* (DfE 2013, p. 13)

This can be compared with the definition of reading given in the PIRLS 2016 reading framework: '*Reading literacy is the ability to understand and use those written language forms required by society and/or valued by the individual. Readers can construct meaning from texts in a variety of forms. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment'*. (Mullis and Martin 2015, p. 12).

The definition taken from the English national curriculum is predictably broader as it encompasses the domains of writing, speaking and listening, and therefore communication, as well as reading. But in both extracts above, it is clear that the skill of reading is seen as providing the learner with a means of participating in society as well as enabling them to learn, and to gain enjoyment, from the act of reading. The national curriculum goes further and recognises reading as contributing to emotional and spiritual growth, rather more ambitious that the reference to 'enjoyment' in the PIRLS' definition. What is interesting is that in the context of what can only be described as the rather 'dry' document that is the statutory national curriculum in England, the moments which lift the text from the arid and extensive lists of what must be taught at certain ages into something that might inspire teachers are when the role of reading in a child's life is described. For example, on page 14, '*Reading also feeds pupils' imagination and opens up a treasure-house of wonder and joy for curious young minds'*. (DfE 2013, p. 14).

Reading as Assessed in England's National Curriculum and in PIRLS

Assessment Frameworks

Test frameworks are published for both the national assessments (DfE 2015a) and PIRLS (Mullis and Martin 2015). Essentially these documents describe the principles behind the assessment – the rationale for the test being designed in the way it is, the balance of content and the skills that it aims to assess. The PIRLS framework summarises the theoretical basis of PIRLS as a measure of reading comprehension. The English national curriculum test framework provides more of a blueprint for test developers – a 'what to do' rather than a justification for the test design. This difference in approach reflects the different roles the assessments have – one provides an international comparative assessment used in over 50 countries, many with very different approaches to the teaching and assessment of reading, and the other a statutory assessment of reading of a national curriculum in one country when the curriculum is already a statutory document.

In this section, these two frameworks are compared, first looking at content, then considering format and test design, and finally looking at the performance standards and demand. Whilst PIRLS assesses the reading skills of grade 4 students (age 9–10), the most comparable national assessments in England are those which are taken at the end of primary school 1 year later, when the students are aged 10–11. For this exercise, the focus is on the main PIRLS assessment, not PIRLS Literacy, designed as a linked assessment but less challenging than PIRLS, nor on ePIRLS, the assessment of digital reading.

Purposes

In PIRLS, two distinct *purposes* for reading are identified: reading for literary experience and reading to acquire and use information. This distinction is followed through in the test design and in some of the subsequent analyses. In the national curriculum test framework, this element is described by a discussion of the *range* of texts to be included in the tests – to include 'fiction, non-fiction and poetry' (DfE 2015a, b). The difference, other than one of vocabulary, is the reference to poetry. A study of poetry is integral to the English national curriculum at all ages; it is, however, inappropriate to include it in an international assessment that will be translated into multiple languages.

Formats

Comparing the *formats* of the two different assessments reveals more similarities than differences. Both use a combination of closed and open response item formats. In PIRLS multiple choice items are the most common closed item format, representing approximately half the marks available. Whilst there are some multiple choice items in England's reading test, closed item formats are a smaller proportion of the items (between 10% and 30%). Open response items require students to construct a written answer. This item type, ranging up to three marks per item but more usually one or two marks, represents around half of PIRLS items and around 80% of England's national curriculum test items.

Demand of the Assessments

As stated above, the assessments are designed for different but adjacent year groups. Taking that 1 year difference into consideration, it is still interesting to compare the descriptions of performance at the standard that could be described as 'proficient', i.e. the point at which students' skills in reading are sufficient to meet the demands of schooling. This is straightforward to identify in England as it is an explicit target in the accountability system – the 'expected standard'. It is less definitive in PIRLS, but elements of the 'Intermediate International Benchmark' and the 'High International Benchmark' appear to describe a set of skills that most closely fit a description of proficiency (see closer Table 9.1.).

The first group of skills included in the English national curriculum and PIRLS Intermediate Benchmark descriptors involves the retrieval of specific details from the text and covers both information/non-fiction and literary/fiction. Clearly the demand of items requiring these skills is affected by the complexity of the text, but these can be among the least demanding items in the tests. The more challenging items assessing this skill provide an inference in the question, and students are expected to identify specific words or phrases from which the inference is drawn.

The second set of skills is those requiring the student to make the inference – this requires the student to create understanding by considering evidence in the text and, in some cases, to explain or justify the inference. This may, for example, be drawing an inference about a character's motivation from their actions or inferring that an animal is dangerous from a description of their physical features. At the PIRLS High Benchmark, the skill is to make global inferences drawing on the whole text. At the Intermediate Benchmark, the skill described is more focused in its demand, often requiring students to refer to a specific section of text (technically known as a local inference).

At this age there is some expectation that students are able recognise language choices. In the national curriculum assessment, it is only occasionally the focus of a question; in the PIRLS Intermediate Benchmark, there is the statement that they

Table 9.1 Comparison of performance standards

England: Expected Standard (age 11)	PIRLS: Intermediate Benchmark (age 10)	PIRLS: High Benchmark (age 10)
Locate information		
Retrieve key details and quotations from fiction and non-fiction to demonstrate understanding of character, events and information	Independently locate, recognise and reproduce explicitly stated actions, events and feelings (L)	Locate and distinguish significant actions and details embedded across the text (L)
	Locate and reproduce two or three pieces of information from text (I)	Locate and distinguish relevant information within a dense text or a complex table (I)
Have an overview of the whole text		
Identify/explain how information in non-fiction is related and contributes to meaning as a whole	Begin to interpret and integrate information to order events (I)	Interpret and integrate story events and character actions, traits and feelings as they develop across the text (L)
Identify/explain how the sequence of events in narrative fiction contributes to meaning as a whole		Integrate textual and visual information to interpret the relationship between ideas (I)
Make accurate and appropriate comparisons within texts		Evaluate and make generalisations about content and textual elements (I)
Accurately and selectively summarise main ideas, events, characters and information in fiction and non-fiction texts		
Make inferences		
Make developed inferences drawing on evidence from the text	Make straightforward inferences about the attributes, feelings and motivations of main characters (L)	Make inferences to explain relationships between intentions, actions, events and feelings and give text-based support (L)
Explain and justify inferences, providing evidence from the text to support reasoning	Interpret obvious reasons and causes, recognise evidence and give examples (L)	Make inferences about logical connections to provide explanations and reasons (I)
Make developed predictions that are securely rooted in the text	Make straightforward inferences to provide factual explanations (I)	
Provide developed explanations for key information and events and for characters' actions and motivations		
Vocabulary		
Show an understanding of the meaning of vocabulary in context		

(continued)

Table 9.1 (continued)

England: Expected Standard (age 11)	PIRLS: Intermediate Benchmark (age 10)	PIRLS: High Benchmark (age 10)
Language use/authorial intent		
Identify/explain how the choice of language enhances the meaning of texts	Begin to recognise language choices (L)	Recognise the use of some language features (e.g. metaphor, tone, imagery) (L)

PIRLS Intermediate International Benchmark (L) = when reading a mix of simpler and relatively complex literary texts, students can …; (I) = when reading a mix of simpler and relatively complex informational texts, students can…
PIRLS High International Benchmark (L) = when reading relatively complex literary texts, students can…; (I) = when reading relatively complex informational texts, students can…

are 'begin[ning] to recognise language choices', whereas in the High Benchmark this is much more ambitious in that the student can 'recognise the use of some language features (e.g. metaphor, tone, imagery)'. Unsurprisingly, given the use of translation, only England's descriptor refers to the explicit understanding of vocabulary although obviously it is implied through the act of comprehension itself.

The main difference between England's 'expected standard' descriptor and the 'Intermediate Benchmark' in PIRLS is the extent to which the reader is expected to consider how specific features contribute to the functioning of the text as a whole – this is not a feature of the Intermediate Benchmark. It is evident in the national curriculum descriptor and is seen more prominently in the High Benchmark on PIRLS in features such as [students can] 'Interpret and integrate story events and character actions, traits, and feelings as they develop across the text' (Mullis et al. 2017).

In summary, there are some similarities in the design and format of the PIRLS and English curriculum assessments described above, in particular in the range of skills the tests endeavour to assess. It is perhaps surprising that there is a greater emphasis on language and style in PIRLS than in the English assessments, given the complexity of assessing in multiple languages, although it is worth noting that this is assessing authorial style and intent rather than vocabulary in context. In terms of the demand, the national curriculum descriptor encompasses elements of both the Intermediate and the High benchmarks, and it is the achievement of these different standards of performance that is discussed in the next section.

Comparing Reading Achievement in England's National Assessments and PIRLS

There is no statistical link between the PIRLS measurement scale and the assessment of reading in the English national curriculum. We cannot know that a score on one equates to a specific score on the other; the tests are developed under different frameworks and also assess students a year apart in age. There are, however, two comparisons that can be made in looking at PIRLS and national assessment data:

Comparing the trends over time in each dataset
Looking at the proportions of students achieving each particular standard on the specific assessment. This section looks at each of these comparisons in turn.

Trends in Achievement

In relation to trends, the outstanding feature of these two measures of reading attainment in England – the national assessments and PIRLS – is that there has been very little change since the first PIRLS survey in 2001. By this time, national assessment was well-established in England and the initial steady annual increase in the proportion of pupils achieving the expected standard had tailed off. Between 2001 and 2006, when the second PIRLS survey was conducted, data shows that the proportion of pupils aged 11 meeting the 'target' standard (level 4) had moved from 82% to 83% (and in the intervening years, had fluctuated between 80% and 84%). Between 2006 and 2011, the second and third PIRLS surveys, the national data shows that again, there was very little change in the proportion meeting the target standard (from 83% to 84%) (DfE 2015b). At the time of the fourth and most recent PIRLS survey in 2016, the new reformed national curriculum was to be assessed for the first time. As, in addition to new content, the 'expected standard' had been explicitly increased, it is not possible to continue the trend discussed above. In 2016, just 66% of pupils achieved the expected standard in reading at age 11 in England. As is expected following the introduction of a new assessment (Ofqual 2016), there was a notable increase in this figure in the two following years (2017: 72%; 2018: 75%) (DfE 2017a, 2018b) although in 2019, a small drop (to 73%) was recorded (DfE 2019).

Proportions Achieving the Standards

The most recent PIRLS data is from 2016. In this survey, 86% of students sampled in England achieved at least the Intermediate Benchmark and 57% achieved at least the High Benchmark. When this cohort was 1 year older, in 2017, 72% achieved the expected standard on the national assessments.

Possible Next Steps in England's Primary Assessment System

The assessment system in England is politically controversial and as a consequence political change is likely to herald changes in statutory assessment. However, the accountability genie is out of the bottle. Formal assessment of student achievement will continue to be used as the leading means of measuring schools' effectiveness in

England. The reception baseline, an on-entry assessment of 4-year-olds, is under development, but how well it meets the objectives in enabling recognition of schools' effective work in the early years of a child's education will be evident only in 2027.

There are, though, other equally fundamental changes to assessment which can be anticipated, although the timescale is uncertain. We can assume that e-assessment will be introduced into national assessments in England in the next decade. Currently the high-stakes nature of testing at key stage 2, when all students in a cohort are assessed on the same day, means that there are very high demands on the schools' infrastructure. This, combined with several high-profile delivery failures, means that there is a very cautious approach to innovation. Scotland and Wales are introducing low-stakes e-assessments, as are some countries in continental Europe. Participation in international surveys of achievement is well-established. The forthcoming delivery of these surveys in an online mode may help pave the way for online statutory assessment. The IEA are refining sophisticated e-assessments for PIRLS and TIMSS, reading in a digital environment as well as reading assessments presented in a linear style on screen. Given the interest in educational measurement, both in terms of national assessments and also the lively interest in international survey results, it seems likely that assessment will remain a topic of some debate in England for some time to come.

References

Buckingham, J. (2016). *Focus on phonics: Why Australia should adopt the year 1 phonics screening check* (Research Report 22). Sydney: The Centre for Independent Studies. https://www.cis.org.au/app/uploads/2016/11/rr22.pdf.

Castle, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest, 19*(1), 5–51. https://doi.org/10.1177/1529100618772271.

Clark, M., & Glazzard, J. (Eds.). (2018). *The phonics screening check 2012–2017: An independent enquiry into the views of Head Teachers, teachers and parents* (Final Report). Birmingham: Newman University. https://www.newman.ac.uk/wp-content/uploads/sites/10/2018/09/The-Phonics-Screening-Check-2012-2017-Final-Report.pdf

Department for Education. (2012). *Assessment framework for the development of the year 1 phonics screening check.* https://www.gov.uk/government/publications/assessment-framework-for-the-development-of-the-year-1-phonics-screening-check

Department for Education. (2013). *The national curriculum in England Key stages 1 and 2 framework document.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/425601/PRIMARY_national_curriculum.pdf

Department for Education. (2015a). *English reading test framework.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/628816/2016_KS2_Englishreading_framework_PDFA_V3.pdf

Department for Education. (2015b). *National curriculum assessments: key stage 2, 2015* (revised). https://www.gov.uk/government/statistics/national-curriculum-assessments-at-key-stage-2-2015-revised

Department for Education. (2017a). *National curriculum assessments: key stage 2, 2017* (revised). https://www.gov.uk/government/statistics/national-curriculum-assessments-key-stage-2-2017-revised

Department for Education. (2017b). *Primary assessment in England government consultation response.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/644871/Primary_assessment_consultation_response.pdf

Department for Education. (2018a). *National curriculum assessments at key stage 1 and phonics screening checks in England (2018).* https://www.gov.uk/government/statistics/phonics-screening-check-and-key-stage-1-assessments-england-2018

Department for Education. (2018b). *National curriculum assessments: key stage 2, 2018* (provisional). https://www.gov.uk/government/statistics/national-curriculum-assessments-key-stage-2-2018-provisional

Department for Education. (2019). *National curriculum assessments: key stage 2, 2019* (provisional). https://www.gov.uk/government/statistics/national-curriculum-assessments-key-stage-2-2019-provisional

Department of Education and Science and Welsh Office. (1988). *National Curriculum: Task Group on Assessment and Testing. A report*. London: DES. http://www.educationengland.org.uk/documents/pdfs/1988-TGAT-report.pdf.

Gibb, N. (2017). *Reading is the key to unlocking human potential.* Speech at the presentation of England's successful Progress in International Reading Literacy Study results at the British Library. https://www.gov.uk/government/speeches/nick-gibb-reading-is-the-key-to-unlocking-human-potential

Hordacre, A. L., Moretti, C., & Spoehr, J. (2017). *Evaluation of the trial of the UK phonics screening check in South Australian Schools*. Adelaide: Australian Industrial Transformation Institute, Flinders University of South Australia. https://www.education.sa.gov.au/sites/g/files/net691/f/evaluation-uk-phonics-screening-check-sa.pdf.

McGrane, J., Stiff, J., Baird, J.-A., Lenkeit, J., & Hopfenbeck, T. (2017). *Progress in International Reading Literacy Study (PIRLS): National Report for England.* https://www.gov.uk/government/publications/pirls-2016-reading-literacy-performance-in-england

Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 Assessment Framework* (2nd ed.). Boston: TIMSS & PIRLS International Study Center, Boston College. http://timssandpirls.bc.edu/pirls2016/framework.html.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 International Results in Reading*. Boston: TIMSS & PIRLS International Study Center, Boston College. http://timssandpirls.bc.edu/pirls2016/international-results/.

National Foundation for Educational Research. (2016a). *NFER education briefings: Key insights from PISA 2015 for the UK nations.* https://www.nfer.ac.uk/media/3049/key_insights_from_pisa_2015_for_the_uk_nations.pdf

National Foundation for Educational Research. (2016b). *NFER education briefings: Twenty years of TIMSS in England.* https://www.nfer.ac.uk/media/1540/99958.pdf

The Office of Qualifications and Examinations Regulation (Ofqual). (2016). *An investigation into the 'Sawtooth Effect' in GCSE and AS/A level assessments.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/549686/an-investigation-into-the-sawtooth-effect-in-gcse-as-and-a-level-assessments.pdf

United Kingdom Literacy Association. (2017). Written evidence submitted by the United Kingdom Literacy Association to the inquiry into primary assessment by the Education Select Committee. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/education-committee/primary-assessment/written/42192.html

Walker, M., Sainsbury, M., Worth, J., Bamforth, H., & Betts, H. (2015). *Phonics screening check evaluation: Final report.* https://www.nfer.ac.uk/publications/yopc03/yopc03.pdf

Whetton, C. (2009). A brief history of a testing time: National curriculum assessment in England 1989–2008. *Educational Research, 51*(2), 137–159. https://doi.org/10.1080/00131880902891222.

Chapter 10
Educational Assessment in Estonia

Gunda Tire

Introduction to International and National Assessment Context and Its History in Estonia

"*Estonian people believe in education. We truly think that the best legacy we can offer to our children is not a piece of land, a house or a bank account, but good education*", said Toomas Hendrik Ilves, the President of Estonia, after learning about the success of Estonian students in PISA 2012. When Estonian media was notified about the release of data from the PISA survey, the press conference room was packed with journalists, waiting for the minister to announce the results. PISA is a well-recognized acronym in Estonian society and the tension before the announcement was high.

Estonia joined PISA in 2006 and its people, being rather modest and critical towards their education system, did not expect anything extraordinary. Results turned out to be most positive, international comparisons ranked Estonia as one of the top performing countries and Estonians experienced a positive "PISA shock". The results were similar to Finland and high-performing Asian countries. Gradually, sceptical Estonians regained confidence in their education system, policy makers got a confirmation that the educational policies and reforms have been successful and schools were reassured that their teaching methods have been effective. PISA data release in Estonia is not only time to actively participate in debates about education but also to recognize and appreciate the job well done by everyone participating in the education system – students, teachers, parents.

High rankings of Estonian students in PISA have brought considerable global attention towards Estonia and its education system. Distinguished education experts, policy makers, researchers, and teachers from all around the world come to Estonia

G. Tire (✉)
Foundation Innove, Tallinn, Estonia
e-mail: gunda.tire@innove.ee

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_10

to see the system first-hand. Their inquiries about the reasons behind the success have made Estonians look in the mirror and reflect about the possible factors that have contributed to the high results. Success in education can be attributed to a multitude of factors, including social, cultural, institutional and historical aspects (Simola 2005). This chapter will attempt to look at some of them.

Estonia and Its School System

Estonia is a small country (45,000 km^2) in Northern Europe, located on the shores of the Baltic Sea, with a population of 1.3 million. It is rich in forests and has nearly 1500 islands along its coastline. The official language is Estonian, which belongs to the Finno-Ugric family of languages. The Estonian population comprises 69% Estonians, 25% Russians and 6% other ethnic groups (Statistics Estonia 2017). The distribution of the population is reflected in the education system. The Estonian education system consists mainly of schools with one of the two languages of instruction – Estonian or Russian.

Estonia is one of the most digitalized societies in the world with numerous online services provided for its citizens. Nearly everything can be done online – filing of tax returns, casting a vote in elections, registering a child birth, etc. The demand for digitally educated citizens puts a significant pressure on the education system (HITSA 2018). Schools have integrated a variety of digital solutions and many teachers use computers, smartboards, robotics kits and other digital devices in their lessons.

Estonia's precarious history has made Estonians adaptive and inventive with regard to different survival strategies, and education has had an important role in the process (Ruus 2002). Formal education dates to the thirteenth century when under the German and Danish rule churches opened the first schools in the territory of Estonia. In the seventeenth century, during the Swedish rule, the first academic schools and the first university were established. In the eighteenth century, the territory of Estonia came under Russian rule, and the education reforms in tsarist Russia applied also to Estonian elementary schools. The first folk schools were opened at the end of the eighteenth century. According to the census in 1897, the level of literacy among Estonians was 79.9% which was the highest in the Russian Empire (Lees 2016). Estonia became an independent state in 1918 and the government introduced general, compulsory and free education for everybody. During the Soviet occupation, education remained in Estonian, strong emphases was put on subjects such as mathematics and science.

A turning point in the history of Estonian education occurred before the break off from the Soviet Union in 1990. In 1987, the Estonian Teachers' Congress criticized the existing school system and demanded an independent Estonian education (Ruus 2002). The first step towards Western education was to develop a completely new, Marxism-Leninism ideology-free curriculum. The development involved participants of different backgrounds and the curriculum was adopted and introduced to

schools before the regaining of independence in 1990. Since 1996, the national curriculum not only states the content of the core subjects, but also emphasizes the need to develop core and cross-curricular competencies. The national curriculum is updated approximately every 10 years, and it states the learning outcomes that students should master during different stages of their education. The legal framework for education in Estonia was established in the 1990s (Lees 2016). The Law on Education (1992) outlines the rights of equal opportunities for everybody. Some significant regulations that followed were the Laws on Basic and Upper Secondary Education Schools (1993), National Curriculum (1996), Law on Private Education (1998).

The Estonian education system is based on a strong pre-school education. Around 94% of children attend kindergartens; children start school at the age of 7. Compulsory education, called the "basic education" lasts from grades 1 to 9. The first streaming into academic or vocational education tracks takes place after grade 9 when students are 15–16 years old. Educational expenses are covered by the state and administration of schools is the responsibility of local municipalities. Schools have considerable autonomy; they develop their own school curriculum, which is based on the framework provided by the national curriculum. Schools have the freedom to decide on the content of optional courses, duration of lessons, and if they opt to specialize in subjects like science, languages, music, art, etc. They can independently choose textbooks and teaching materials, principals can hire and fire teachers, decide on school resource allocation and plan teacher training (Innove 2016).

The education system was "upgraded" in 2014, when the Estonian government adopted "the Estonian Lifelong Learning Strategy 2020". The document provides guidelines for strategic development in education and serves as the base for funding decisions in education. The document foresees five priorities for development:

1. Change in the approach to learning, orientation to progressive and student-centred approaches
2. Empower competent and motivated teachers and school leadership with knowledge on modern approaches and practices
3. Align lifelong learning opportunities with needs of the labour market
4. Digital focus in lifelong learning
5. Equal opportunities and increased participation in the lifelong learning. (Ministry of Education and Research 2014a, b)

All goals are elaborated in detail; they are linked with indicators from national and international assessments and are annually measured.

National and International Assessments

As schools are autonomous, and education is financed by the taxpayer money, certain accountancy mechanisms are required for steering and measuring the efficiency of the system. The state has the right to get feedback on how well students have

mastered the educational goals set in the national curriculum (Ministry of Education and Research 2014a, b). With that in mind, the external evaluation system was established in the mid-1990s. Its main components are sample based tests for grades 3 and 6, as well as national examinations for grades 9 and 12. Centrally provided tests for grades 3 and 6 are low stakes tests, as students are not individually graded, and are hugely popular among schools. Many non-sampled schools administer the tests to their students as schools consider them to be a valuable feedback reflecting the instructional quality of the school. Data is shared with parents and other stakeholders. Grade 6 tests are all computer based. The strategic goal is to have computer-based national assessment system up and running by 2021 as intended in the Estonian Lifelong Learning Strategy 2020.

At the end of compulsory education, in grade 9, students take three centralized exams, whereas the marking takes place at school by the subject teacher. The requirements for finishing the basic school consist of centralized examinations in Mathematics, Estonian language, one freely chosen subject by the student (from a list of 10 subjects), and a completed research project organized by the school.

At the end of upper secondary school, grade 12 students take three centrally set and centrally marked national examinations that are valid also for entering universities or other higher educational establishments. Students should pass the national examinations in Estonian or Estonian as a second language, mathematics (two different curricula are offered with different number of learning hours and corresponding exams), and in a foreign language. In addition to centralized examinations, students are required to pass a school exam, and conduct an independent research project in the topic of their interest.

The first international student assessment where Estonia participated was IEA study TIMSS 2003 (Trends in International Mathematics and Science Study). That was followed by OECD PISA (Programme for International Student Assessment) 2006, 2009, 2012, 2015 and 2018. Estonian teachers and principals have participated in OECD survey TALIS (Teaching and Learning International Survey) in 2008, 2013 and 2018. Student readiness to be future citizens has been assessed by IEA International Civic and Citizenship Education studies in 2009 and 2016. Estonia has also participated in the OECD PIAAC (Programme for the International Assessment of Adult Competencies), which is a survey on adult skills in literacy, numeracy, and problem solving in technology rich environment. The international assessments have provided a full picture of Estonian education from student, teacher and system points of view.

Findings of International Assessments

According to OECD data, Japan, Estonia, Finland and Canada are regarded as the four highest-performing OECD countries (Organisation for Economic Co-operation and Development 2016a, b, c). The domain in which Estonian students have excelled in all international assessments is science. Estonian students scored 534 points in

science in PISA 2015, exceeded only by Singapore (556 points) and Japan (538 points). High performance in science was already noted in TIMSS 2003 – the first international student assessment survey where Estonia participated. The unexpectedly high results were, subsequently, repeated in the following PISA cycles. Results have been stable in science, slight improvements observed in mathematics and a larger increase in reading literacy.

According to PISA data, the Estonian education system is not only high performing but also ranks high in equity. Students from different socio-economic backgrounds have good access to education, and they achieve high results. Only 8% of the score variance in science is explained by students' socio-economic background. As much as 48% of students are the so-called resilient students, which is the sixth highest result among participating countries. Resilient students come from the bottom quarter of the PISA index of economic, cultural social status and perform among the top quarter of students among all countries after accounting for socio-economic status (Organisation for Economic Co-operation and Development 2016a, b, c).

Another high point of Estonian education system is that it has a relatively small share of students who perform below the baseline level of proficiency. PISA highlights student performance not only according to the mean scores but also by the distribution of scores on levels of proficiency (levels 1–6). The higher the level, the more complex are the tasks the student can solve, level two being the baseline level of proficiency. Only 8.8% of Estonian students score below level two; this share is smaller only in Vietnam and Macao (China). Baseline knowledge in science has been reached by 91.2% of Estonian students (OECD mean 78.8%). There is no performance gap between boys and girls in science; also the share of top performers has increased since 2006 (Innove 2016). The need to pay more attention to high-performing students has been a priority in the educational discourse and the effect of efforts is seen in the more recent data of the international assessments. Many schools have started to pay more attention to students who could reach higher levels of proficiency, and organize different activities to develop their full potential.

In examining Estonian policy documents, we notice that the high equity, as seen in PISA, complies with the principles of the comprehensive school, rooted in the legislation in the 1990s. The policy includes equity and inclusiveness. For example, all students get free school meals, free textbooks, access to different extra-curricular activities, free school transport, etc. Schools must provide the best learning environment for everyone, regardless of students' family background. If needed, students should receive additional instruction and have access to services of psychologists, social pedagogues, speech therapists or other support. Grade repetition is rather exceptional; students should get help on time to move on (Innove 2016). The system cares for the weakest students, and the small share of low-performing students in PISA reflects that.

An important lesson Estonia learned from the international assessments concerns the performance difference of schools with Estonian and Russian language of instruction. Although the Russian-medium schools have improved over time, the performance gap is still equal to approximately one school year of learning, which

on the PISA scale is about 39 points. All schools get the same funding; have the same guidelines from the national curriculum and conditions for learning, etc. Additional research has been done and it shows that the gap in science performance could be explained by mainly two factors: (1) the socio-economic status of the household and parents' educational level, and (2) student attitudes and beliefs (enjoyment of studying science and epistemological beliefs) (Täht et al. 2018).

International assessments shed light not only on the student cognitive outcomes but also on their background information. Issues of student well-being, learning environment, learning habits have been significant to learn about. The Estonian Lifelong Learning Strategy 2020 has determined that student well-being is an important aspect in education.

The PISA 2015 survey asked students how satisfied they were with their life in general at the time of the test. The scale of possible responses ranged from 1 to 10. The higher the number, the higher life satisfaction students reported. The mean score for Estonian students was 7.6, which shows a rather high level of students' life satisfaction. OECD has ranked Estonia (together with Finland, the Netherlands, and Switzerland) as one of the countries with high student performance and high life satisfaction. Collected background information also points out areas which need a constant check, like bullying, sense of belonging, student truancy, and so on (Organisation for Economic Co-operation and Development 2017a, b).

Already in 2002, the Estonian national curriculum introduced problem-solving skills, and social and emotional skills as important components of the education outcome. Mastering of these competencies should be integrated into the teaching process and not taught as separate subjects. In PISA 2015, the innovative domain was collaborative problem solving. This was a good chance to see how well Estonian students could solve problems with interactive, "virtual" companions in unfamiliar situations. Estonian students scored sixth among the participating countries with a mean score of 535 (exceeded only by Singapore (561), Japan (552), Hong Kong (541), Korea (538) and Canada (535)). This confirmed that teaching of the "soft skills" like teamwork is applied effectively in Estonian schools.

The international survey that has given voice to schools to speak about their experiences is called TALIS (Teaching and Learning International Survey). TALIS studies teachers and school principals from around 200 schools per country, explores issues about initial teacher training and continuous professional development, providing feedback, school and classroom climate, etc. It also asks the teachers how satisfied they are with their job and how they feel about their profession (Organisation for Economic Cooperation and Development 2014). Estonia has participated in TALIS three times (2009, 2013 and 2018).

What is the portrait of an average Estonian teacher according to TALIS 2013? Teachers are mostly female (84.5%); the average age is 48; average teaching experience is 21.6 years; 95% of teachers have the required qualification and 35% of teachers, mostly older teachers, work part time. This is not always their own choice as in smaller schools there are not enough lessons or students, to be employed full time. At the same time there is an overall shortage of teachers, especially in science subjects. Estonian schools and class sizes are rather small; on average, there are

17.3 students per class. The school year in Estonia is among the shortest in world with 175 days, and teachers can enjoy 2 months of summer holidays. Although school principals think that the school climate is positive, problems listed concern mostly mental bullying, truancy, student cheating. Especially less experienced teachers feel that they are not valued enough by society. However, they all like their job and the school environment (Organisation for Economic Cooperation and Development 2014).

TALIS has pointed out a problem that the teacher population in Estonia is aging and there is an increasing shortage of teachers. Teacher salaries have been considerably increased during the last years; however, they are still low compared to absolute numbers of other OECD countries (Organisation for Economic Cooperation and Development 2018).

What have been the contributors to Estonian student success? As the distinguished American education expert Marc Tucker mentioned after his visit to Estonia in his blog "*it is this combination of low pay, the small number of days in the school year, the high workload for teachers and high student performance that makes Estonia's system so efficient*" (Tucker 2015). This observation points to serious issues in the sustainability of the system's effectiveness, and similar concerns have been expressed by the OECD observers (Organisation for Economic Cooperation and Development 2018).

Different activities are done at the state level to promote the image of the teacher's profession. A well-rooted tradition in Estonia is the celebration of Teacher's day at the beginning of October. To identify and reward the best teachers in the country, a national nomination and award ceremony is organized. The event is called "Estonia Learns and Thanks" where teachers in different categories from all over the country are nominated and awarded (Ministry of Education and Research 2018a, b). The award ceremony is aired on the national television; it always draws a big audience and is later energetically debated in the media. There are also many programmes provided by the government and co-financed by the European Union (ESF) that aim to promote professionalism of teachers and school leaders.

How to Go Further? Where Next?

International assessments have shown that the education system in Estonia is high performing and effective; however, it has some sustainability issues concerning the future of the teaching profession. Education is a process in progress and policy decisions about the future of the education system is, in large part, a political process (Weiss 2001). Education policy makers recognize that they are influenced by factors such as scientific studies, organizations, people and information sources (Swanson and Barlage 2006). This suggests that international studies such as PISA and TALIS contribute partially to the process of educational policy making.

The main components of the national external evaluation system are the assessment of learning outcomes, and the evaluation of schools; however, guidelines for

the future directions for Estonian education come from the Estonian Lifelong Learning Strategy 2020. Five goals in the strategy document have specific indicators to be reached and the goals are supported with substantial funding by the state. The first goal in the Estonian Lifelong Learning strategy 2020 moves the education away from the traditional ways of teaching towards a progressivist, child-centred educational approach. In this "changed approach to learning", personal and social development of each learner should be encouraged, as well as the development of their learning to learn skills, fostering creativity and entrepreneurship during all levels and types of education (Ministry of Education and Research 2014a, b). In regard to the future of educational assessment in Estonia, the strategy document shifts the attention towards formative assessment, which should support learning and the individual development of each learner. The focus is on the learners, their key skills and cross-curricular competences.

According to legislation, the goal of external assessment is to give students, parents, schools, school administrators and the state an objective and comparative feedback to the learning objectives stated in the national curriculum, as well as provide an input for education policy making (Põhikooli- ja gümnaasiumiseadus 2010). Considering the legislation and strategy document, the Ministry of Education and Research has published a plan for 2020. It focuses on the following ideas:

1. Support every student, teacher, school

This focus complies with the goal to enhance digital technologies in teaching and learning. The Ministry of Education and Research has launched the development of innovative digital assessment models. A lot of effort is put into the development of computer-based "diagnostic tests" that would detect what students already know and what are their gaps in a specific topic or skill. Literature suggests (Christodoulou and William 2017) that when students learn new material, 75% of them make the same mistakes, but for different reasons. Teachers, at the same time, often do not have a good overview of the topics students have mastered well or those they have failed to understand. The solution to the problem is a feedback system or diagnostic tests that find evidence about student learning. Computer-delivered diagnostic tests decrease the teacher's workload and provide teachers with immediate feedback, for example, about the effectiveness of the teaching methodology applied in the teaching process. Based on the test results, the teacher can quickly determine what material has been mastered by the students and what needs more attention. As a result, teaching and learning becomes more effective. Diagnostic tests are currently in the development phase. Other sets of computer-based tests, developed in collaboration with universities, are tests in digital literacy and tests in certain key skills like learning to learn.

2. Collect supportive evidence to decide about development of students and schools

Until recently, the evidence collected about student learning outcomes was the information from the sample-based centrally set subject tests in grades 3 and 6, and final exams for grades 9 and 12. A recent addition to the external evaluation is the study on student well-being.

The Estonian Lifelong Learning Strategy 2020 sets a goal to increase student well-being, improve school learning environment and increase participation rates in lifelong learning. In 2015, a decision was made to create an instrument to measure the progress towards the set goals and a theoretical framework of measurement was developed. Pilot studies took place in 2016 and 2017 and the first full-scale well-being survey was administered in 2018 to all students from general education schools in grades 4, 8 and 11. To make the picture more complete, separate questionnaires for teachers and parents of the participating students were used. In addition, pre-primary and vocational education establishments were included in the well-being survey. The goal is to get the big picture at the system level and to provide individual schools with comparative reports about the general well-being and school climate. Each school gets a detailed report that provides indicators about general well-being of students, teachers and parents of the school. It also points out the problematic areas for the school to work on and improve (Ministry of Education and Research 2019). Schools use this data as an input for evidence-based self-development and quality improvement. The creation of centralized well-being measurement tool has spared schools from developing their own questionnaires on this matter and added quality and comparability to the whole system.

3. Make suggestions for education decision making at the state level

External evaluation provides feedback about the implementation of national curriculum. It also suggests where changes should be made in the national curriculum, in teachers' continuous professional development, in textbooks and in teaching and learning process.

4. Inspire schools in the learning process

External motivation is needed for generation of internal motivation and for encouragement of autonomous motivation (Ryan and Deci 1999). It is expected that the new, state-developed computer-based tests and digital learning materials will assist teachers to get fast feedback about student achievement, detect knowledge gaps and adjust their teaching accordingly. The on-demand digital tools enable constant monitoring of students' errors and can be used as a rich source for finding next learning assignments and filling previous knowledge gaps. Digital tests come along with digital item banks in different subjects that enable teachers to individualize learning and group students for different activities (Innove 2019).

As already mentioned, this is work in progress. The effectiveness and the school approval of the new digital tools are yet to be seen. Due to the autonomous school system, where schools and teachers can choose how they teach their students, the adoption of the new tools for learning and its measurement to some extent is based on voluntary principles. There is a strong push from the policy makers to use digital tools in the learning and assessment phases. Digital solutions with all the interactive possibilities engage children very strongly and allow studying anywhere (Kikas 2018).

As the literature suggests, education policy making to a large extent is a political process (Swanson and Barlage 2006). At the beginning of academic year 2018–2019,

the Minister of Education and Research announced that external evaluation in Estonia should move away from the culture of testing and lean towards individual student support. International assessments have shown that in comparison to other countries, Estonia is rather modestly subjected to standardized testing (Organisation for Economic Cooperation and Development 2017a, b). The minister's proposal was to abolish the centralized final examinations at the end of grade 9, which has been a significant component of external evaluation, and let schools themselves decide if their students have completed the compulsory education. Another suggestion was to extend the list of national exams for grade 12, which was not a new idea. Until 2012, the list of exams provided by the state consisted of 14 exams, which was substantially cut down by a different decision maker.

Since effective education is a balance between rigour and freedom, tradition and innovation, the individual and the group, theory and practice (Robinson and Aronica 2016), there is a hope that the autonomous Estonian schools will assimilate the suggested innovations in a good balance with the best educational traditions.

Well, what is the secret of Estonian success? Marc Tucker put it this way: "*The fact that Estonia is among the top performers in PISA does not appear to be the result of education policies pursued since Estonia gained its independence, but rather the result of hundreds of years of political, social and educational development which ended up supporting a strong commitment to education as well as a tradition of very high education standards, very demanding curriculum, high quality examinations built directly on the curriculum, highly educated teachers, and most of the other drivers of high performing national education systems*" (Tucker 2015).

References

Christodoulou, D., & William, D. (2017). *Making good progress? The future of assessment for learning*. Oxford: Oxford University Press.

Estonian Ministry of Education and Research. (2014a). *The Estonian lifelong learning strategy 2020.* Retrieved from https://www.hm.ee/en/estonian-lifelong-learning-strategy-2020

Estonian Ministry of Education and Research. (2014b). *Üldhariduse välishindamise ülesanded, põhimõtted ja arendamise alused aastani 2020*. Retrieved from https://www.hm.ee/sites/default/files/uldhariduse_valishindamise_ulesanded.pdf

Estonian Ministry of Education and Research. (2018a). *Eestimaa õpib ja tänab.* Retrieved from https://www.hm.ee/et/gala

Estonian Ministry of Education and Research. (2018b). Mailis Reps: tähelepanu all on alusharidus, digiõpikud ja õpetajaameti väärtustamine. Retrieved from https://www.hm.ee/et/uudised/mailis-reps-tahelepanu-all-alusharidus-digiopikud-ja-opetajaameti-vaartustamine

Estonian Ministry of Education and Research. (2019). *Rahulolu haridusega.* Retrieved from https://www.hm.ee/et/rahulolu

Foundation Innove. (2016). *Estonia in the spotlight: PISA 2015.* Retrieved from https://www.innove.ee/wp-content/uploads/2018/03/PISA_ENG_2015_voldik_web_final.pdf

Foundation Innove. (2019). *Digiõpevara.* Retrieved from https://www.innove.ee/oppevara-ja-metoodikad/digioppevara/e-ulesandekogud/diagnostilised-testid/

HITSA strategy for 2018–2020. Retrieved from https://www.hitsa.ee/about-us/news/hitsa-strategy-for-2018-2020-supports-increase-in-digital-competencies-in-the-society-as-a-whole

Kikas, E. (2018). *E-ülesanded ja Diagnostilised Testid.* Retrieved from https://youtu.be/jUoJl0LC90M

Lees, M. (2016). *Estonian education system 1990–2016: Reforms and their impact*. Retrieved from http://4liberty.eu/wp-content/uploads/2016/08/Estonian-Education-System_1990-2016.pdf

McIntyre, N. A. (2014). *Increasing achievement in science education: learning lessons from Finland & Estonia.* Retrieved from https://www.wcmt.org.uk/sites/default/files/report-documents/McIntyre%20N%20Report%202014.pdf

Organisation for Economic Cooperation and Development. (2001). Reviews of National Policies for Education: Estonia 2001, Reviews of National Policies for Education. https://doi.org/10.1787/9789264189966-en.

Organisation for Economic Cooperation and Development. (2014). *TALIS 2013 results: An international perspective on teaching and learning.* https://doi.org/10.1787/9789264196261-en.

Organisation for Economic Cooperation and Development. (2016a). *PISA 2015 high performers: Estonia.* Retrieved from https://www.oecd.org/pisa/PISA-2015-estonia.pdf

Organisation for Economic Cooperation and Development. (2016b). *PISA 2015 results (Volume I): Excellence and equity in education.* https://doi.org/10.1787/9789264266490-en.

Organisation for Economic Cooperation and Development. (2016c). *PISA 2015 results (Volume II): Policies and practices for successful schools*. https://doi.org/10.1787/9789264267510-en.

Organisation for Economic Cooperation and Development. (2017a). *PISA 2015 results (Volume III): Students' well-being*. https://doi.org/10.1787/9789264273856-en.

Organisation for Economic Cooperation and Development. (2017b). *PISA 2015 results (Volume V): Collaborative problem solving.* https://doi.org/10.1787/9789264285521-en.

Organisation for Economic Cooperation and Development. (2018). *Education at a glance 2018: OECD indicators*. https://doi.org/10.1787/eag-2018-en.

Põhikooli- ja gümnaasiumiseadus. (2010). Retrieved from https://www.riigiteataja.ee/akt/13332410

Robinson, K., & Aronica, L. (2016). *Creative schools*. London: Penguin Books.

Ruus, V. (2002). *The history of Estonian education – the story of the intellectual liberation of a nation.* Retrieved from http://www.estonica.org/en/Education_and_science/The_history_of_Estonian_education_%E2%80%94_the_story_of_the_intellectual_liberation_of_a_nation/

Ryan, R. M. & Deci, E. L. (1999). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology 25*, 54–67 (2000). Retrieved from https://mmrg.pbworks.com/f/Ryan,+Deci+00.pdf

Santiago, P., et al. (2016). *OECD reviews of school resources: Estonia 2016*, OECD reviews of school resources. https://doi.org/10.1787/9789264251731-en.

Simola, H. (2005). The Finnish miracle of PISA: Historical and sociological remarks on teaching and teacher education. *Comparative Education, 41*(4), 445–470.

Statistics Estonia. (2017). *Minifacts about Estonia 2017.* Retrieved from https://www.stat.ee/publication-2017_minifacts-about-estonia-2017

Swanson, C. B., & Barlage, J. (2006). *Influence. A study of the factors shaping education policy. Editorial projects in education research centre.* Retrieved from https://www.edweek.org/media/influence_study.pdf

Täht, K., Konstabel, K., Kask, K., Rannikmäe, M., Rozgonjuk, D., & Schults, A. (2018). *Eesti ja vene õppekeelega koolide 15-aastaste õpilaste teadmiste ja oskuste erinevuse põhjuste analüüs.* Retrieved from https://www.hm.ee/sites/default/files/uuringud/pisa_ev_raport_0507_006.pdf

Tucker, M. (2015). *Tucker's lens: Estonia: Unsung Heroine of the Baltic, but…* Retrieved from http://ncee.org/2015/03/tuckers-lens-estonia-unsung-heroine-of-the-baltic-but/

Weiss, C. (2001). *What kind of evidence in evidence based policy?* Retrieved from https://www.cem.org/attachments/ebe/P284-291%20Carol%20Weiss.pdf

Chapter 11
Educational Assessment in Finland

Mari-Pauliina Vainikainen and Heidi Harju-Luukkainen

The Finnish education has received a lot of attention after decades of relatively high performance in international student assessments. Even though the Finnish education system has received a lot of interest, very little attention has been paid to the model of the Finnish educational assessment system and the lack of standardised measurement and control. Thus, these factors in large are contributing to the overall functioning of the system. In this chapter we provide a historical overview of the development of the assessment model in Finland and further give a description of its current form. We also give an example of the Finnish PISA 2012 oversampling and its results. Finally, we make some critical suggestions on how the system could be improved without adding unnecessary controlling elements to it.

Introduction

The Finnish education system, still after almost two decades, continues to receive international attention due to the relatively high performance level of its students in international educational assessment studies like the Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) (e.g. OECD 2016). The education system in Finland differs from the high-performing countries in several aspects. An important difference is that the results of Finland are high despite average economic investments into education (OECD 2018).

M.-P. Vainikainen (✉)
Tampere University, Tampere, Finland
e-mail: mari-pauliina.vainikainen@tuni.fi

H. Harju-Luukkainen
Nord University, Levanger, Norway

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_11

Finnish schools also seem to be exceptionally equitable in terms of the low level of segregation both by the distribution of socio-economic status of pupils and by their performance levels (Willms 2010). Also the school level differences are among the smallest ones in the world (OECD 2016).

When it comes to the basic principle of the education system, there are a lot of commonalities between the different Nordic countries (Garvis and Eriksen Ødegaard 2017). Therefore, Finland is implementing alongside with the rest of the Nordic countries a so-called Nordic model (Antikainen 2006; Telhaug et al. 2006) or sometimes even defined as a 'Nordic dialogue' (Garvis and Eriksen Ødegaard 2017). According to this model, all students should have equal opportunities and possibilities regardless of their socio-economic status or residential area. Children are also attending their local school, without any tracking, until they are at the end of their compulsory education. This model was introduced in Finland in the Basic Education Act of 1968, and it was gradually implemented from 1972 to 1976.

In Finland, children start their school the year they turn seven. In most cases, children are enrolled in their local school, but in some cases, there is a possibility to apply to a school with a specialised language, music or other programmes during the lower grades. The emphasis on children attending the local public school was strengthened in the legislation reform in 2011 (Thuneberg et al. 2013), and the statistics from the following year show that 96% of the comprehensive schools were run by municipalities (the Official Statistics of Finland, www.stat.fi). Therefore, there are almost no private actors among the Finnish schools. Some adjustments have been made to the education system since the 1970s. For instance, the ability grouping widely practiced during the first decade was officially abolished in the mid-1980s, but otherwise the structure remains unchanged. Different to many other countries, the first official point of tracking occurs only after the ninth year of schooling. At this point, students have to choose between academic and vocational tracks of upper secondary school.

Even though the Finnish education system has received a lot of interest, very little attention has been paid to the model of the Finnish educational assessment system and the lack of standardised measurement and control (see Vainikainen et al. 2017). Thus, these factors in large are contributing to the overall functioning of the system. The aim of this chapter is firstly to provide a historical overview of the development of the assessment model in Finland and further to give a description of its current form. We also give an example of the Finnish PISA 2012 oversampling and its results. Finally, we make some critical suggestions on how the system could be improved without adding unnecessary controlling elements to it. After all, the freedom of municipalities, schools and teachers in organising the education according to their best understanding and implementing different aspects of the curricula might have a higher role in shaping the good educational outcomes of the Finnish youth than previously recognised.

Introduction to the National and International Assessment Context and Its History in Finland

The Finnish educational assessment model has evolved in several stages until it has reached its current, relatively noncontrolling structure (Varjo et al. 2016). The early decades of the comprehensive education system were characterised by a strict control of particularly *inputs* and to some extent also *outputs* (cf. OECD 2015). The inputs were regulated through a detailed national curriculum with state-level obligatory in-service teacher training of obligatory contents of it and pre-examination of textbooks. Outputs were controlled by an active school inspection system that held schools accountable for achievement, the same way it is nowadays done in many other countries (Gustafsson et al. 2015). However, the attempt of introducing standardised national exams in major school subjects in the 1970s failed, and therefore teacher-given school grades became the primary measure of achievement. At that time, grades followed a normal distribution within each class, leading to a situation, in which between-school or class differences were not recognised. Also standardised test was provided for teachers only to facilitate the grading process.

During curriculum reform in 1985, the obligatory national curriculum was replaced by a *National Framework Curriculum*, which gave the municipalities as organisers of education more freedom for local decision-making. It also gave them wider possibilities to assess the outputs of the system locally even though at this point, the national school inspection system was still active. Besides that, there was no national assessment system, and the country's participation in the International Educational Assessment (IEA) surveys was irregular. As the recommendations given in the National Framework Curriculum on grading were relatively unspecific, too, grading was still largely done at class level on a normative scale even though the aim was to adopt a criterion-referenced model and not to rank students.

The next curriculum reform took place in 1994, taking local-level decision-making on the next stage. The *national core curriculum* contained only the obligatory core, and municipalities and/or schools had to write their own curricula based on it. Organisers were also expected to assess the outcomes of the education they provided. Further, also the school inspection system and controlling of the learning materials had been ceased before the new core curriculum was introduced. Therefore, at this time Finland experienced its first period of decentralisation of monitoring mechanisms for educational outputs, even though practices for evaluating educational outcomes (see National Board of Education 1999, for the revised English version) had not yet been formally introduced. This, however, did not last long as the final version of the assessment practices for evaluating educational outcomes was published in 1995 and the more comprehensive version in 1998 simultaneously with the educational legislation reform. Educational assessment was now also defined as the organiser's responsibility in the legislation. In practice, these local assessments were largely based on self-evaluation.

The national monitoring model for evaluating educational outcomes was introduced (National Board of Education 1999). It was designed to provide some indicators of performance trends and further needs for the national policy development work. On an international scale, this model was light, and accountability was not a part of it on any level. On the contrary, the introduced approach made it almost impossible to make any conclusions on individual school level, and this new type of data was solely for national monitoring purposes.

In this model the educational outcomes, were divided into three main categories that all comprised several subcategories. The first category, *efficiency*, measured the functioning of the educational system, whereas the second category *effectiveness* was about student-level outcomes. The third category was *economy* for successful allocation of resources. From the perspective of educational assessment, the second category was the most interesting one as it posed requirements of conducting external assessments. In practice, this led to two kinds of applications. First, sample-based *curricular assessments* were developed to measure learning outcomes in most important school subjects. Second, *national thematic assessments* were developed to provide information on a wider scope of educational outcomes to complement the information obtained from international large-scale assessments that were now emerging: as the first cycle of PISA in 2000 and by rejoining the IEA assessments. Curricular sample-based assessments are still in 2018 not implemented regularly at predefined grade levels in major subjects, but the subjects and grade levels to be assessed are instead specified for a few years at a time in a *plan for educational assessment*. The current organisation of these assessments is described below. In the 1990s, the National Board of Education that was also responsible for curriculum development had a unit for implementing them.

In the mid-2000s, Finland had the second period without a clear national structure for educational assessment. For the whole decade, the National Board of Education implemented curricular sample-based assessments, and Finland participated in PISA (and more irregularly in TIMMS and PIRLS, but not between 1999 and 2011). However, the national coordination of assessments was restructured several times (see Varjo et al. 2016), which led to a situation where many thematic assessments – including the national learning to learn assessment programme – were ceased even though no formal decisions were made of not having them. At the same time, the new core curriculum of 2004 continued to give organisers of education a lot of freedom and responsibility to define their own assessment practices to fulfil the requirements stated in the 1998 legislation. Both general and subject-specific assessment criteria were specified and harmonised with curricular goals, and descriptions of 'good performance' were given to facilitate both formative and summative assessments. Themes introduced during the previous decades about cross-curricular or transversal competences were to some extent included in the core curriculum as general goals, but relatively little was said about them in assessment criteria or descriptions about subject-specific goals.

National and International Assessment in Finland Today

National Assessments

Until recently, the Finnish National Board of Education used to conduct national assessments on students' learning achievement. Since May 2014, this is now a duty of the Finnish Education Evaluation Centre (FINEEC). Among other things, the FINEEC is responsible for evaluating learning outcomes with respect to the distribution of lesson hours and the national core curriculum targets stipulated in the Basic Education Act (628/1998). The assessment of learning outcomes is based on sampling. Typical sample sizes comprise 5–10% of the age group, which means that each assessment involves about 4000–6000 students (Jakku-Sihvonen 2013, 24). The assessed schools represent around 15% of all the schools that give basic education in Finland (Ouakrim-Soivio, 2013, 21; Harju-Luukkainen et al. 2016a).

According to Harju-Luukkainen et al. (2016a), the assessment of learning outcomes can be viewed from many perspectives, and it has got different purposes for different target groups. National assessments provide valuable information for the highest educational authorities. In Finland, basic education is expected to secure equal educational opportunities for all students. Therefore, the equity of learning outcomes is studied from several perspectives, for example, those of students' gender, region, type of municipality and socio-economic background as well as language spoken at school. In principle, reaching the objectives for equal learning opportunities as defined in the national core curricula should lead to educational equity so that there would be no statistically significant differences between the learning outcomes of boys and girls, for example, or between different regions in Finland.

Secondly, from the school's perspective, the national assessments of learning outcomes provide benchmarks for schools to evaluate their own success in reaching their objectives of teaching and learning in different subjects. Schools selected to an assessment receive feedback in the form of reference data on the results and learning-related perceptions of their own students. Because there are no national examinations at the end of basic education, many schools welcome this opportunity to compare their own results and grading practices to the national benchmarks and use the assessment as a tool to develop their instruction in different subjects (Ouakrim-Soivio and Kuusela 2012, 13; Harju-Luukkainen et al. 2016a).

Thirdly, teachers assess each student based on student performance. At the end of basic education (i.e. grade 9 in the comprehensive school), most of the students are 15-year-olds and about to finish their compulsory education. Grading is obligatory at the final phases of basic education (grades 8 & 9), but most schools begin to use numerical grades already at earlier grade levels. In Finland, the national core curriculum for basic education determines the learning objectives for each school subject. Also grading guidelines are given but with specific description for good

competence only, which equals the grade 8 on the student assessment scale ranging from 4 to 10, where 4 means failed and 10 is the highest grade. This good competence level serves as a baseline for assessment, and it should help ensure an objective evaluation for all students attending basic education. Objective evaluation at this point is of great importance; the grades obtained in different subjects at the end of compulsory education will largely determine the next steps in the student's educational path. In sum in Finland today, student assessments at different levels (national, school or individual level) all strive for the same goal: higher equality in education (Harju-Luukkainen et al. 2016a).

International Assessments

Finland has been participating in PISA assessments since the first cycle in 2000. In addition, Finland has taken part to IEA assessments (PIRLS, TIMSS) in the recent cycles, but earlier the country's participation in these assessments has been too irregular to monitor trends based on these data. Thus, at basic education level, national discussions about performance trends in international assessments are largely based on PISA. When the results of the PISA 2000 cycle were first published, they provoked surprisingly little public discussion. The high ranking was not expected, and the reception of the news was almost sceptical. Yet, it most likely changed the course of educational political discussion as there had been voices claiming that the comprehensive education system does not support the optimal development of students with higher academic goals. The results showed that high-performing students did not do any worse than their counterparts in other countries, whereas the weakest students clearly outperformed their comparison groups anywhere else (Hautamäki et al. 2009). Thus, the results may have contributed to the basic education legislation changes that have strengthened the main principles of the Nordic educational ideas even further (see Thuneberg et al. 2013). The first PISA results were used as evidence for that the structure of the education system did not need extensive reforms. Accordingly, the declining trend observed both in international and national assessment studies since 2006 (Hautamäki et al. 2013; Vettenranta et al. 2016) has been taken seriously, and programmes have been launched to turn the trend again. These programmes have included thematic assessments that go deeper into the details of the national features of the education system (e.g. support structures) and additional funding for municipalities and schools to improve their practices. The data of international assessments are also utilised in this purpose through more detailed analyses.

An Example: Oversampling PISA 2012 in Finland

Like many other European societies, Finland has faced many changes during the past decade. One major change has been the increasing number of students speaking languages other than the national ones. Due to the increase in the Finnish migrant population, students with a migrant background were oversampled for the first time in PISA 2012. In this, Finland was the second country to conduct an oversampling of one of its student population, after Denmark. Oversampling means that more students were selected for testing than would be their true proportion in the population. The oversampling made it possible to gain more representative data on students with migrant backgrounds. The Finnish PISA data on migrant students consisted of 691 first-generation and 603 second-generation students, most of whom lived in the metropolitan areas. The rest of the data comprised a total of 7535 students across the country (see Harju-Luukkainen et al. 2014). In the following, we will give an overview of these results (see more Harju-Luukkainen and McElvany 2018).

A first report on these national findings was published in 2014. The PISA 2012 migrant data was analysed and the reports published just before the migration crisis hit Europe in 2015. During this time, more than a million migrants and refugees crossed Europe, coming mainly from Syria, Afghanistan, Iraq, Kosovo and Albania to Finland among other countries. Therefore, the oversampling of students with a migrant background in Finland gave an important insight on how Finland had managed in educating their language minority students.

Even though Finland is known for its good educational outcome, the results were not that positive. According to Harju-Luukkainen et al. (2014), the results of students with migrant backgrounds were alarming compared to their nonmigrant peers in Finland. As shown in Fig. 11.1, students with migrant backgrounds performed poorer in PISA 2012 across all assessment domains compared to their nonmigrant peers. The differences were statistically significant. In mathematics, for instance, nonmigrant students received a mean score of 522 points; first- and second-generation students with migrant backgrounds received 425 points and 449 points, respectively. The definition of who is counted as first- and second-generation migrants can vary in different studies. However, according to the OECD and PISA 2012 assessment, first-generation migrant students are those who have immigrated to Finland during their lifetime. Second-generation students in turn have been born and raised in Finland, but both of the parents were born outside Finland. According to the OECD (2014, p. 16), 41 points scored equals approximately 1 year of schooling. First-generation students were therefore lagging behind by more than 2 school years and second-generation students by almost 2 school years. The results were more or less similar in scientific literacy and reading literacy as well. Similar

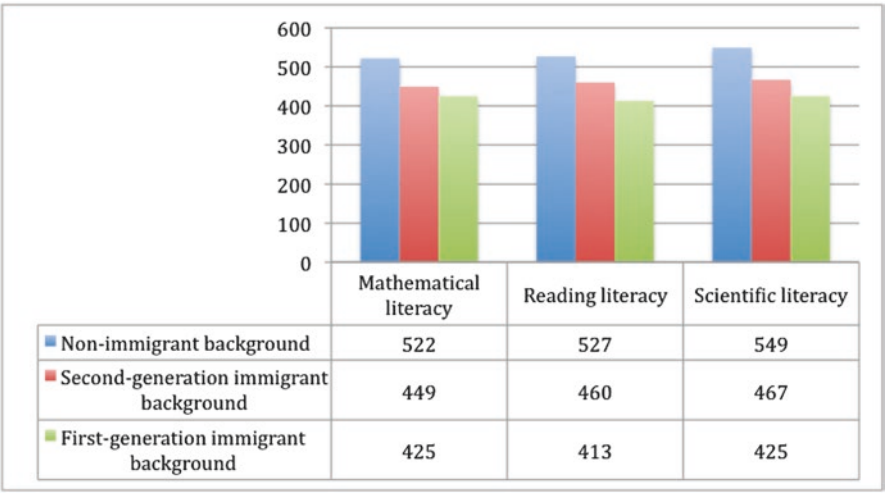

Fig. 11.1 Mathematic, reading and scientific literacy of different student groups in PISA 2012. (Source: Harju-Luukkainen et al. 2014, p. 25)

Table 11.1 Percentage of students on different performance levels in PISA 2012 (Harju-Luukkainen et al. 2014, p. 27)

Performance level	First generation	Second generation	Nonmigrant
6	0.7	0.4	3.6
5	2.3	2.4	12
4	8.3	10.4	23.7
3	14.9	22.1	29.2
2	22.2	26.6	20.4
1	26.2	24.2	8.3
Below level 1	25.3	13.9	2.7
	100%	**100%**	**100%**

observations have been done in all other PISA cycles even though the migrant sample has been smaller.

In the data especially the low percentage of high performers as well as high level of low performers among the migrant students was observed. According to Harju-Luukkainen et al. (2014) on the highest performance levels (levels 5 and 6), there were almost no students with a migrant background (varying between 0.4% and 2.4%) and a very small difference between the proportions of first- and second-generation students at the highest performance levels. Further, 51.5% of first-generation migrant students were at the lowest performance levels (level 1 and below), as were 38.1% second-generation migrants. According to Harju-Luukkainen and McElvany (2018), it is troubling when the second-generation migrants have taken part of the entire Finnish education system, the performance is still on a very

low level, and the difference between the first- and second-generation migrants is relatively small (Table 11.1).

The migration population in Finland is still relatively small and heterogenous. For this reason, there have only been a few studies looking into the reasons behind these differences. According to Harju-Luukkainen et al. (2014), mathematical literacy performance in all native student groups in Finland was explained by such variables as self-concept for mathematics, confidence in mathematics performance and anxiety for mathematics (see also Harju-Luukkainen et al. 2016b). However, the explanatory power of these variables was weaker for students with a migrant background than for other students. Therefore, there is most likely a wider range of underlying factors for this minority group that are either unknown or at least beyond the scope of PISA assessments. Harju-Luukkainen et al.'s (2015) investigation of resilient second-generation migrants students' educational outcomes in mathematics found that the factors connected to good educational outcomes were (1) the family's language choices, (2) high ESCS (student's socio-economic and cultural status index), (3) cultural closeness, (4) teacher's support and individualisation of teaching materials, (5) low truancy and intact learning continuums and (6) strong self-concept in mathematics (see also Harju-Luukkainen et al. 2017). How well a student masters the language of instruction seems therefore to be one of the most important factors, which is something that Kuukka and Metsämuuronen (2016) and Saario (2012) also emphasise. The study conducted by the FEEC (Kuukka and Metsämuuronen 2016) revealed that migrant pupils' Finnish language skills were good, already in the upper grades of comprehensive school. However, the concept of text skills of various subjects requires more from the pupils than is required by the criterion of a proficiency scale. Therefore, it is crucial to ask if the different assessments capture the true level of migrant students' competencies and skills.

The degree to which these family-related attributes have an impact on students' educational outcomes varies not only from country to country (OECD 2010) but also within countries (Harju-Luukkainen and McElvany 2018). In Finland, there are to be found differences between the different student groups and how their family-related attributes affect the students' educational outcome. According to PISA 2012, the ESCS index (student's socio-economic and cultural status index) explained 11% of the variance between first-generation students, 7% of the variance between second-generation students and 8% of the variance between students without migrant backgrounds (Harju-Luukkainen et al. 2014). Further, the assessment conducted by the FEEC revealed that variables which could best explain students' low learning outcomes from different language groups were related to students' socio-economic background. In all, the connection between the socio-economic background and learning outcomes was significant (Kuukka and Metsämuuronen 2016). In these abovementioned studies (as well as many other studies), the ESCS has not been controlled. In a study conducted by Kilpi-Jakonen (2012, p. 167), the study revealed that differences between migrant and nonmigrant student groups in Finland are relatively small after controlling for parental resources. Kilpi-Jakonen (2011; 2012) concluded that parental education and parental income have smaller and larger effects, respectively, for children of migrants than for nonmigrants. This

leads to a disadvantaged group with migrant parents who have high education levels but low incomes (Harju-Luukkainen and McElvany 2018). According to Kalalahti et al. (2017), the youth with migrant background in Finland, especially boys, share a 'paradox of immigrant schooling' which refers to the positive attitude towards education, but at the same time, they face difficulties in learning and studying. Overall, according to Kilpi-Jakonen (2012), children of migrants can be seen to benefit from the relatively equal Finnish education system while remaining disadvantaged by their parents' difficulties in the labour market (see more Harju-Luukkainen and McElvany 2018; Karppinen 2008; Kilpi-Jakonen 2011).

Critical Discussion of the Country's Assessment Policies, Practices and Results

Finland has developed its assessment policies and streamlined many of the practices during the history of its basic education. A further developmental object is to find a balance in how the results of the national assessments as well international assessments are used on national level.

However, still a national central organ that oversees both the national and international assessment policies, practices as well as results does not exist. This has lead in some cases to a situation, where participation in some important international assessment has been irregular and there has been problems with the capability of national sample-based assessments to produce enough comparable data for monitoring of trends. In the Finnish context, it still seems functional not to have extensive standardised examinations as the lack of them gives more freedom to schools and teachers to implement the curriculum in a purposeful way. However, the declining trend and the increasing regional differences (Vettenranta et al. 2016) call for a slightly more detailed monitoring system that could be realised within the current educational assessment model by securing sufficient coverage of school subjects, age groups, at-risk subpopulations and geographical areas.

Lack of resources is also a risk for Finnish assessment context. The government of Finland should direct enough of funding towards the analysis of already collected datasets in order to reveal, for instance, possible negative trajectories behind educational outcomes of different student groups. As the situation is now, only basic reporting and data collection can be done with the governmental funding. This might lead to a situation that education policy decisions, which are solely done on the basis of national reports, can be weakly justified.

References

Antikainen, A. (2006). In search of the Nordic model in education. *Scandinavian Journal of Educational Research, 50*(*3*), 229–243.

Garvis, S., & Eriksen Ødegaard, E. (2017). *Nordic dialogues on children and families. Evolving families*. Oxon: Routldege.

Gustafsson, J. E., Ehren, M. C. M., Conyngham, G., McNamara, G., Altrichter, H., & O'Hara, J. (2015). From inspection to quality: Ways in which school inspection influences change in schools. *Studies in Educational Evaluation, 47*, 47–57.

Harju-Luukkainen, H., & McElvany, N. (2018). Immigrant student achievement and educational policy in Finland. In L. Volante, D. Klinger, & Ö. Bilgili (Eds.), *Immigrant student achievement and education policy: Cross-cultural approaches* (pp. 87–102). Cham: Springer.

Harju-Luukkainen, H., Nissinen, K., Sulkunen, S., & Suni, M. (2014). *Avaimet osaamiseen ja tulevaisuuteen: Selvitys maahanmuuttajataustaisten nuorten osaamisen tasosta ja siihen liittyvistä taustatekijöistä PISA 2012 – tutkimuksessa* [Keys to competence and future: A report on PISA 2012 results and related underlying factors for students with an immigrant background]. Jyväskylä: Finnish Institute for Educational Research.

Harju-Luukkainen, H., Nissinen, K., & Mirja, T. (2015). Matematiikka ja maahanmuuttajataustaiset nuoret [Mathematic literacy and immigrant students]. In J. Välijärvi & P. Kupari (Eds.), *Millä eväillä tulevaisuuden peruskoulu nousuun?* (pp. 108–123). Helsinki: Ministry of Education and Culture.

Harju-Luukkainen, H., Vettenranta, J., Oukrim-Soivio, N., & Bernelius, V. (2016a). Differences between PISA reading literacy scores and grading for mother tongue and literature at school: A geostatistical analysis of the Finnish PISA 2009 data. *Education Inquiry, 7*(4), 463–479. https://doi.org/10.3402/edui.v7.29413.

Harju-Luukkainen, H., Sulkunen, S., & Stolt, S. (2016b). Uppfattning av lärandestrategier hos elever som för det mesta talar finska hemma, men går i svenskspråkig skola [Understanding of reading related strategies among students who speak mostly Finnish at home but attend a Swedish language school]. *Kognition och Pädagogik, 99*, 78–89.

Harju-Luukkainen, H. Tarnanen, M., & Nissinen, K. (2017). Monikieliset oppilaat koulussa: eri kieliryhmien sisäinen ja ulkoinen motivaatio sekä sen yhteys matematiikan osaamiseen PISA 2012–tutkimuksessa [Multilingual students in school: The inner and outer motivation of students with an immigrant background and its connection to mathematical performance in PISA 2012]. In A. Huhta & R. Hildén (Eds.), *Kielitaidon arviointitutkimus 2000-luvun suomessa* [Language assessments in 20th century Finland] (pp. 167–183). Jyväskylä: Suomen soveltavan kielitieteen yhdistys AFinLA ry.

Hautamäki, J., Hautamäki, A., & Kupiainen, S. (2009). Educational equity account in Nordic countries. In T. Matti (Ed.), *Northern lights on PISA 2006: Differences and similarities in the Nordic countries* (TemaNord 2009:547) (pp. 157–167). Copenhagen: Nordic Council of Ministers.

Hautamäki, J., Kupiainen, S., Marjanen, J., Vainikainen, M. –P., & Hotulainen, R. (2013). *Oppimaan oppiminen peruskoulun päättövaiheessa: Tilanne vuonna 2012 ja muutos vuodesta 2001* [Learning to learn at the end of basic education: Situation in 2012 and change from 2001] (University of Helsinki. Department of Teacher Education Research Report 347). Helsinki: Unigrafia.

Jakku-Sihvonen, R. (2013). Oppimistulosten arviointijärjestelmistä ja niiden kehittämishaasteista [On assessment systems for learning outcomes and developmental challenges of these systems]. In Räisänen, A. (Ed.). *Oppimisen arvioinnin kontekstit ja käytännöt* [Contexts and practices for the assessment of learning] (Raportit ja selvitykset 2013:3, pp. 13–36). Helsinki: Opetushallitus.

Kalalahti, M., Varjo, J., & Jahnukainen, M. (2017). Immigrant-origin youth and the indecisiveness of choice for upper secondary education in Finland. *Journal of Youth Studies., 20*, 1242. https://doi.org/10.1080/13676261.2017.1321108.

Karppinen, K. (2008). Koulumenestys, koulutukseen valikoituminen, tutkinnon suorittaminen ja työelämään siirtyminen [School success, selection of education, graduation and transition to work]. In J. Kuusela, A. Etelälahti, Å. Hagman, R. Hievanen, K. Karppinen, L. Nissila, U. Rönnberg, & M. Siniharju (Eds.), *Maahanmuuttajaoppilaat ja koulutus Tutkimus oppimistuloksista, koulutusvalinnoista ja työllistämisestä* (pp. 135–186). Helsinki: Opetushallitus.

Kilpi-Jakonen, E. (2011). Continuation to upper secondary education in Finland: Children of immigrants and the majority compared. *Acta Sociologica, 54*(1), 77–106. https://doi.org/10.1177/0001699310392604.

Kilpi-Jakonen, E. (2012). Does Finnish educational equality extend to children of immigrants? Examining national origin, gender and the relative importance of parental resources. *Nordic Journal of Migration Research, 2*(2), 167–181. https://doi.org/10.2478/v10202-011-0039-4.

Kuukka, K., & Metsämuuronen, J. (2016). *Perusopetuksen päättövaiheesn suomi toisena kielenä (S2) –oppimäärän oppimistulosten arviointi 2015* [Estimating the learning outcomes of Finnish second language (S2)]. Tampere: Kansallisen koulutuksen arviointikeskus.

National Board of Education. (1999). *A framework for evaluating educational outcomes in Finland* (National Board of Education, Evaluation 8/1999). Helsinki: University Printing House.

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

OECD. (2014). *PISA 2012 results: What students know and can do – Student performance in mathematics, reading and science* (Vol. I). Paris: OECD.

OECD. (2015). *Education at a glance 2015* (OECD Indicators). Paris: OECD Publishing. https://doi.org/10.1787/eag-2015-en.

OECD. (2016). *PISA 2015 results (volume I): Excellence and equity in education*. Paris: OECD Publishing. https://doi.org/10.1787/9789264266490-en.

OECD. (2018). *Education at a glance 2018* (OECD Indicators). https://doi.org/10.1787/eag-2018-en.

Ouakrim-Soivio, N. (2013). *Toimivatko päättöarvioinnin kriteerit? Oppilaiden saamat arvosanat ja Opetushallituksen oppimistulosten seuranta-arviointi koulujen välisten osaamiserojen mittareina* [Do the national criteria for students' final grades work? Teachers' grading and national assessments as indicators for between-school differences]. Helsinki: Opetushallitus.

Ouakrim-Soivio, N. & Kuusela, J. (2012). *Historian ja yhteiskuntaopin oppimistulokset perusopetuksen päättövaiheessa 2011* [Learning outcomes in history and social studies at the end of basic education 2011] (Koulutuksen seurantaraportit 2012:3). Helsinki: Opetushallitus.

Saario, J. (2012). *Yhteiskuntaopin kieliympäristö ja käsitteet. Toisella kielellä opiskelevan haasteet ja tuen tarpeet* [The language environment and concepts of social work: Challenges and support needs in another language] (Doctoral dissertation). University of Jyväskylä, Finland.

Telhaug, A. O., Mediås, O. A., & Aasen, P. (2006). The Nordic model in education: Education as part of the political system in the last 50 years. *Scandinavian Journal of Educational Research, 50*(*3*), 245–283.

Thuneberg, H., Vainikainen, M.-P., Ahtiainen, R., Lintuvuori, M., Salo, K., & Hautamäki, J. (2013). Education is special for all – The Finnish support model. *Gemeinsam leben, 21*(*2*), 67–78.

Vainikainen, M.-P., Thuneberg, H., Marjanen, J., Hautamäki, J., Kupiainen, S., & Hotulainen, R. (2017). How do Finns know? Educational monitoring without inspection and standard-setting. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard setting: International state of research and practices in the Nordic countries* (pp. 243–259). Cham: Springer. https://doi.org/10.1007/978-3-319-50856-6_14.

Varjo, J., Simola, H., & Rinne, R. (2016). *Arvioida ja hallita. Perään katsomisesta informaatio-ohjaukseen suomalaisessa koulupolitiikassa* [To evaluate and govern: From "looking after" to steering by information" in Finnish education policy] (Publications of the Finnish Educational Research Association, 70). Jyväskylä: Jyväskylän yliopistopaino.

Vettenranta, J., Välijärvi, J., Ahonen, A., Hautamäki, J., Hiltunen, J., Leino, K., Lähteinen, S., Nissinen, K., Nissinen, V. Puhakka, E., Rautopuro, J. & Vainikainen, M. -P. (2016). *PISA 2015 ensituloksia. Huipulla pudotuksesta huolimatta.* Opetus- ja kulttuuriministeriön julkaisuja 2016:41.

Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teachers College Record, 112*(*4*), 1007–1037.

Chapter 12
Educational Assessment in Germany

Nele McElvany and Justine Stang

Germany participates in several national (e.g. NEPS) and international (e.g. PISA, TIMSS) student assessments. This chapter thus presents an overview of Germany's participation in several national and international assessments. It also explains Germany's current educational monitoring system and how it has been affected by results of international student assessments. Results from the Progress of International Reading Literacy Study (PIRLS) are described in detail as an example of international student assessments in Germany. The study focuses on students' reading literacy, students' motivation, instructional quality, differences between boys and girls, and differences between students with and without a migration background. The chapter closes by discussing Germany's assessment policies, practices and outcomes.

Introduction

The Education System in Germany

In Germany, the responsibility and cultural sovereignty for the education system lies primarily with the 16 federal states. Children, mainly aged three to six, may attend kindergarten. Subsequently, children are enrolled in school. In most states, the primary school lasts for four years. Therefore, the educational system is selective at a very early age: By the end of primary school, the decision, mainly based on students' achievement, is made to which type of secondary school children may go. The secondary schools are separated in lower- and upper-secondary schools (e.g.

N. McElvany (✉) · J. Stang
Center for Research on Education and School Development (IFS), TU Dortmund University, Dortmund, Germany
e-mail: nele.mcelvany@tu-dortmund.de

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_12

grammar school). The first phase of secondary education typically lasts for 6 school years (grade 5–10), leading to several options of school leaving certificates. However, the length of compulsory education differs between the federal states. After compulsory schooling, another decision on the educational path has to be made. At this stage, the German education system is characterized by a wide range of education and training tracks, including, for example, upper-secondary school, which leads on to A levels or vocational schools. While 41.2% of all students graduated school with a school leaving certificate allowing them to study at a German university in 2016, about 6% left school without any formal graduation certificate (Autorengruppe Bildungsberichterstattung 2018).

History of Participation in International Assessments in Germany

Although (West) Germany started to participate in international student assessments early on, their scope was limited. In 1964, two German federal states took part in the First International Mathematics Study (FIMS) that assessed mathematics performance in secondary school students, mathematics teaching and the influence of social, curricular, and technological developments across 12 countries (Husén 1967; Schultze and Riemenschneider 1967). Shortly after FIMS, Germany joined parts of the Six Subject Study (English; political education); and in 1971, a representative sample of students from ten federal states (Schultze 1975) took part in the First International Science Study (FISS). These studies were conducted by the International Association for the Evaluation of Educational Achievement (IEA), founded in 1958 as an international cooperative of national research institutions, governmental research agencies, scholars and analysts. In 1971, the intergovernmental economic organization, the Organisation for Economic Co-operation and Development (OECD) also published an examination of educational systems in several countries including Germany.

After this early phase of Germany's (at least partial) interest in international educational comparisons, there was a long interval before Germany participated in an international large-scale study again with a representative student sample. In 1990–1991, 9- and 14-year-old Germans from both East and West German federal states were included in the International Study of Reading Literacy (IRLS or RL; Lehmann et al. 1995). Although the performance of German students was only just average, these findings attracted limited attention. In the same period, 10- and 13-year-olds were tested in 9 federal states in the 1989 and 1992 Computers in Education Study (ComPed; Lang and Schulz-Zander 1994; followed later by the Second Information Technology in Education Study). Participation in these large-scale assessments (LSA) marked the beginning of a new phase of educational monitoring that shifted from a nearly two-decade-long focus on issues such as individual school development or school tracking. The educational administrations' approach

on input orientation in these decades was met by an educational science often emphasizing on other approaches than empirical-quantitative evidence. This new phase continued with the German participation of representative samples of 13-year-olds and young adults in the Third International Mathematics and Science Study (TIMSS) from 1994 to 1996. The merely average performance outcomes raised awareness of the need for both empirical assessments of educational outcomes and improvements in teaching math and science in German schools. For education practice, one outcome was a large-scale model programme to 'Increase of the efficiency of the math- and science instruction' (SINUS) initially just for secondary schools but later also for primary schools ending in a transfer project period (Prenzel et al. 2009). For educational research, this led to the development of a strong area of research on math and science education that advanced instructional research compared to most other subjects by both quantity as well as empirical foundation. At the end of the millennium, Germany joined the Civic Education Study (CIVED) of representative samples of 14-year-olds in 28 countries (Händle et al. 1999). Again, although German students' performance was only average, the results received little attention in Germany from either the general public or educational administrators. In 2009, Germany opted out of the study, and in 2016, only one federal German state joined the International Civic and Citizenship Education Study (ICCS).

The awareness of all stakeholders – educational policymakers, administrators, practitioners, universities educating future teachers, educational researchers, and last, but not least, the German public – was raised suddenly and definitively by the results of the Programme for International Student Assessment (PISA) study (Baumert et al. 2001; OECD 2001). The study revealed that 15-year-olds in Germany were performing far below expected levels in reading literacy (484 points and thus 16 points below the international OECD average), and that the distance to the group of high-performing countries was substantial. Additionally, the subgroup of very low performers was large (about 20%), as was the performance heterogeneity within Germany, and performance was particularly bad in the 'reflecting and judging' subscale. It also turned out that performance correlated more closely with socioeconomic family background than in any other country under investigation. In the same vein, the average performance of students with a migrant background was much worse than that of native students. One year later, the results of the Progress in International Reading Literacy Study (PIRLS) were published and showed a more positive picture of German primary school students' reading literacy than the PISA results for secondary school students (Bos et al. 2003; Martin et al. 2003): the average performance of German 4th-graders was the same as that in other participating European countries, the variation in performance was comparatively small (indicating homogeneity), and girls and boys did not differ as much in their reading skills as they did in many other countries.

Germany continued to participate in the regular waves of TIMSS (only 4th graders), PIRLS and PISA in the following years (see Table X.1). It also took part in the IEA's International Computer and Information Literacy Study (ICILS) in 2013 and 2018. Furthermore, Germany also joined (1) the OECD LSA Programme for the

International Assessment of Adult Competencies (PIAAC) focusing on adult competencies in literacy, numeracy and problem solving within technology-rich environments (2011); (2) the Teacher Education and Development Study in Mathematics (TEDS-M; 2008; financed by a grant from the German Research Foundation to Humboldt Universität Berlin) focusing on future math teachers; and, currently, (3) the Teaching and Learning International Survey (TALIS-Video; 2018; financed by a grant from the Leibniz Society).

Educational Monitoring and Policy Documents, Perspectives and Assessment Strategies in Germany

After a long phase of abstinence from international educational studies and disappointing results in newer LSA, the politics of educational monitoring in Germany were significantly changed towards a systematic overall approach that was supported by all 16 federal states in the last 20 years. This has also included a fundamental shift from a previously long-term emphasis on input to a new orientation towards output in the perceptions and measures of educational administrators and state governments. An integral part of this new approach has been the establishment in 2004 of an academic institute funded by all the German federal states: the Institute for Educational Quality Improvement [Institut zur Qualitätsentwicklung im Bildungswesen, IQB]. The institute's purpose is to ensure and improve the quality of education by operationalizing and evaluating educational standards and coordinating standard-based item development (see also Klieme et al. 2004). The change on the policymaking level has been accompanied by a substantial transformation of German educational science from the earlier domination of non-(quantitative) empirical approaches towards a strong empirical foundation in much current research. This development gained momentum through the establishment of a broad number of professorships dedicated to empirical educational research at most German universities and the subsequent formation of research groups in this field. Recent German assessment strategies based on the adjustment from an input to an output orientation are presented in three core documents agreed upon by the Standing Conference of the Ministers of Education and Cultural Affairs (KMK 1997, 2006, 2016), which includes ministries from all 16 federal states:

1997: Educational policymakers declared their aim to use empirical data from educational research to identify strengths and weaknesses in the educational system and to use appropriate measures; focus on secondary schools (Grades 9/10) and competencies in (German) language, mathematics, science, and foreign languages; and personal and social skills. Known as the '*Konstanzer Beschluss'*, this marked the start of the shift towards an empirical approach. [Empirische Wende]

2006: Core studies and instruments were defined as a shared basis for an evidence-based educational policy oriented on the results (output) of educational processes (known as the '*Plöner Beschlüsse*').

2015: Update to the overall strategy, strengthening the need for explanatory next to descriptive knowledge, and identifying core areas of interest for further evidence to guide educational policy and practice.

All in all, four areas or tools have been identified and agreed upon as the current focus of educational monitoring in Germany:

1. Participation in international large-scale assessments (PIRLS, PISA, TIMSS)
2. Evaluation and implementation of educational standards [*Bildungsstandards*]: national assessments that enable comparisons across federal states and evaluate whether students meet the educational standards defined for specific subjects in specific grades; focus on the end of primary, secondary and continued secondary education with centralized tests at the end of the first two phases and the provision of a central pool of tasks for the final examination qualifying students for university entrance
3. Quality assurance on the school level: state-specific and cross-state implementation of assessments, making it possible to compare the performance of individual schools and classes in order to support instruction and school development (e.g. 'VERA' [*Vergleichsarbeiten*])
4. Bi-annual publication of a comprehensive national report on the status of educational system by the Federal Ministry of Education (BMBF) together with all German states

The six thematic areas of particular interest identified by the federal states in 2016 (KMK 2016) are:

(1) Heterogeneity: individual support in heterogeneous learning groups including special needs and gifted students
(2) Development of instruction: effects of instructional methods and didactic concepts, usage of evidence-based measures to ensure quality of instruction and school development
(3) Relevance of teacher education and teacher deployment for students' academic development
(4) Effects of measures of school quality assurance
(5) All-day schools: consequences for learning outcomes
(6) Effects and strategies of school development: differences between schools in similar settings

International and National Assessments Today

Today, Germany still participates regularly in core international large-scale student assessments and has joined some additional studies (see Table 12.1). In most cases, study implementation in Germany is commissioned by the Federal Ministry of Education and Research (BMBF) and/or the Standing Conference of the Ministers

Table 12.1 Germany's current participation in multi-wave international large-scale assessments

Study	Short	International research coordination	National research coordination	Rhythm	Last assessment (and previous assessments)	Next assess-ment	Main assessment focus	Main target group	Number of participants
Progress in International Reading Literacy Study	PIRLS	IEA	TU Dortmund University	Every 5 years	2016 (2001, 2006, 2011)	2021	Reading	4th graders	~4000 students in ~200 classes
Trends in Mathematics and Science Study	TIMSS	IEA	University of Hamburg	Every 4 years	2015 (1995, 2007, 2011)	2019	Mathematics, science	4th graders	~4000 students in ~200 classes
Programme for International Student Assessment	PISA	OECD	Centre for International Student Assessment (ZIB)	Every 3 years	2018 (2000, 2003, 2006, 2009, 2012, 2015)	2021	Reading, mathematics, science	15-year-olds	~ 10,000 students in ~230 schools
International Computer and Information Literacy Study	ICILS	IEA	Paderborn University	Every 5 years	2018 (2013)	2023	Computer and information literacy	8th graders	~3000 students in ~150 schools
Programme for the International Assessment of Adult Competencies	PIAAC	OECD	GESIS Leibniz Institute for the Social Sciences	Every 10 years	2011/12	2021	Literacy, numeracy, problem solving in technology-rich environments	16- to 65-year-olds	~5000 persons

Note: IEA International Association for the Evaluation of Educational Achievement, *OECD* Organisation for Economic Co-operation and Development

of Education and Cultural Affairs (KMK). The national research coordinators for the different studies are located at various German universities.

Current national assessments in Germany include the IQB studies evaluating with representative samples from all federal states whether students meet set educational standards. Every 5 years, 4th graders are assessed in German language and mathematics (2011, 2016; upcoming: 2021); and every 3 years, 9th graders are assessed either in mathematics and science (2012, 2018; upcoming: 2024) or in German language or second language English/French (2009, 2015; upcoming: 2021). Yearly assessments of all students in 3rd and 8th grade – in at least German language and mathematics – are the responsibility of the individual federal states and serve the different purpose of helping teachers and administrators to further develop instruction and schools. Focusing on research evidence, Germany also started the National Education Panel Study (NEPS) in 2010 (Blossfeld et al. 2011). This multi-cohort longitudinal study is investigating how education develops from early childhood to old age and the effects education has on other aspects of life. Two starting cohorts (SC 2 and 3) began to follow students from the beginning of primary respectively secondary school education, whereas another starting cohort (SC 4) began with 9th-grade students. Since 2006, the status of the German educational system from early child care institutions, schools, professional education, university education up to adult further training is being reviewed every second year by a national education report commissioned by federal and state educational ministries (Klieme et al. 2003; Autorengruppe Bildungsberichterstattung 2018). These official measures are accompanied by a multitude of additional studies and reports (e.g. the yearly expert report by Aktionsrat Bildung 2018).

An Example of International Assessment Findings: PIRLS

Reading literacy, the focus of PIRLS, is essential for succeeding in academic, working, and everyday life (McElvany et al. 2008). Reading literacy includes the ability to extract relevant information from texts and to understand, use and reflect on written texts (Mullis and Martin 2015). Several national and international studies have shown repeatedly that a substantial number of students have deficits in reading comprehension at the end of primary school (Baumert et al. 2001), thereby indicating the importance of measuring and monitoring students' learning in reading.

PIRLS monitors trends in the reading literacy of 4th graders. Since 2001, PIRLS has been administered every 5 years. Every student reads two texts, one literary and one informational, and works on 12–15 comprehension questions. In addition, students answer questions on their motivation, their attitudes towards reading, and their perception of instructional quality. Furthermore, parents, teachers and school principals complete questionnaires gathering information on students' reading comprehension and the school and family background.

Germany has taken part in every cycle since 2001, and will also participate in 2021 by surveying approximately 4000 4th-grade students from about 200 primary

schools in all 16 federal states. Germany's participation is part of the overall educational monitoring strategy agreed upon mutually and funded equally by the KMK and the BMBF. In Germany, PIRLS is being coordinated by the Center for Research on Education and School Development (IFS) at the TU Dortmund University.

In 2016, Germany's 4th-grade students scored an average of 537 points (Bos et al. 2017). Compared to other countries, Germany ranked in the lower middle range. Nonetheless, this mean score of 537 points is significantly higher than the international average (521 points) and is not significantly lower than the mean score for EU countries (540 points) or all OECD countries (541 points). In 2016, Germany was outperformed by 14 participating states (e.g. Sweden, Italy, Australia) and one benchmark state. In the long-term perspective from 2001 to 2016, there has been no significant change in German students' reading achievement (2001: 539, 2006: 548, 2011: 541 and 2016: 537 points). This result is comparable to other countries such as Sweden, Denmark, or Bulgaria. Even though there was no significant increase in students' reading achievement, the proportion of students on the highest competence level has increased (Competence level V: from 8.6% in 2001 to 11.1% in 2016). Parallel to this, however, there has also been an increase in low-level readers (under competence level III: from 16.9% in 2001 to 18.9% in 2016; Bos et al. 2017). This indicates that the heterogeneity of achievement has become larger. Indeed, students' achievement variance (78 points) was very high in 2016 and comparable to that in countries such as Hungary and Lithuania (Bos et al. 2017).

PIRLS differentiates between literary and informational reading literacy. In Germany, 4th-grade students had higher scores on literary reading literacy (542 points) than on informational reading literacy (533 points). This finding is comparable to 19 other participating countries, although the difference between literary and informational reading literacy is exceptionally large in Germany (Bos et al. 2017).

As in all other countries except Macao SAR and Portugal, girls in Germany (2016) showed a higher average achievement than boys did. Girls outperformed boys especially in literary reading literacy (18 points). The difference was notably smaller (5 points) for informational reading literacy. In 2001, the difference between girls and boys was almost the same.

Fourth-grade students from families with more than 100 books in the home had a 54-point achievement advantage over students with fewer books in the home. This result is similar for all participating states. Alongside Slovenia, Slovakia and Hungary, Germany is one of the four states in which social disparities have increased significantly since 2001. Furthermore, there is also a migration-related gap in reading achievement. Students who did not speak the test language at home had a lower average score than native speakers. In Germany, this difference amounts to 37 points. This means in effect that these students had a disadvantage of 1 year of learning. Students whose parents were born in Germany scored an average of 48 points more than students whose parents were born in a foreign country. All in all, achievement disparities between 4th graders with and without a migration background have remained constant over the last 15 years (Bos et al. 2017).

Related to student achievement is student motivation. In Germany, most 4th graders were highly motivated (2016: $M = 3.18$, $SE = 0.03$ on a 4-point scale ranging from 1 [*disagree a lot*] to 4 [(*agree a lot*]). However, from 2001 to 2016, there has been a decrease in reading motivation, especially in students with reading difficulties. There were also gaps between girls and boys as well as between students with and without a migration background: girls had a higher reading motivation than boys did, and students without a migration background were more motivated than those with a migration background. Additionally, approximately 70% of students read once or twice a week in their free time. Again, girls read more than boys, and students without a migration background read more than those with a migration background (Bos et al. 2017).

To some extent, student achievement and student motivation are an outcome of instructional quality (Hattie 2008). Instructional quality can be divided into three major domains (Hamre and Pianta 2010): classroom management, cognitive activation and emotional support. Classroom management was perceived as very efficient by 39% of students; 60% of those students who rated classroom management as being efficient belonged to the group of high achievers. For the cognitive activation domain, 57% of students felt that they receive strong cognitive activation from their teacher. Most of these students (49%) were high achievers. The last domain, emotional support from the teacher, rated most positive: nearly 73% of students perceived that they got a strong emotional support from their teacher. Of this group, 55% were high achievers (Stahns et al. 2017).

The increasing heterogeneity of classrooms makes it necessary to determine which factors are most relevant for the acquisition of reading literacy under such changing conditions. Both the quantity and the quality of PIRLS data make it possible to examine these questions in detail. Drawing on PIRLS 2016, Hartwig et al. (submitted) combined student and teacher data to analyze whether the interplay between student heterogeneity in cognitive abilities and teachers' attitudes towards heterogeneity such as perceived costs and utility or instructional behavior (especially differentiated instruction) influence students' reading literacy. Hartwig et al. (submitted) found positive relations between teachers' perceived utility and students' reading literacy as well as negative correlations between differentiated instruction and reading literacy on the classroom level. In addition, they found that students' heterogeneity in cognitive abilities related positively to their reading achievement. After controlling for the mean cognitive abilities in classes, only the path between teachers' differentiated instruction and students' reading literacy remained statistically significant (Hartwig et al. submitted).

Germany benefits greatly from participation in PIRLS. PIRLS empirical data have been used to initiate different policies to increase equity in educational opportunity (Wendt et al. 2017). Based on the results of PIRLS 2001, 2006, 2011, and 2016, several measures should be implemented to establish equal opportunities independent of students' gender or migration background, to promote reading also beyond primary school, or to give equal support to both low and high achievers (Valtin 2017).

The fifth PIRLS cycle in 2021 will provide data spanning two decades. Additionally, PIRLS 2021 will offer several new initiatives. Through the transition to a digital format, PIRLS will also be assessing informational and literary reading digitally. In addition, ePIRLS, which was initiated in 2016, will measure students' online informational reading competencies. Another reform will be the national school panel: about 120 schools that already participated in PIRLS 2016 will be retested. This national school panel can be used to analyze longitudinal processes as well as developments on a school level. In Germany, PIRLS 2021 will focus on topics such as instructional quality while taking account of digitalization, multi-criteria goal attainment, and current topics such as the integration of refugee students and mainstream inclusion.

Critical Discussion

Germany's performance in international and national assessments varies depending on the age group and the domain under investigation. Nevertheless, there are strong overall signs of a failure to achieve satisfactory success on such important goals as (a) increasing average results, (b) enlarging the high-performing group, (c) reducing the low-performing group, (d) reducing the correlation between socio-economic family background and performance level, and (e) providing more effective support for students with a migrant family background. The findings from international large-scale assessments have had and continue to have a substantial impact on many levels of education in Germany. Core consequences are the shift from an input to an output orientation and the subsequent continuous evaluation of the educational system and outcomes based on quantitative empirical data. These adjustments in approach have had comparable effects on educational policies, educational administration, educational practice, and even educational science. The 16 federal states of Germany have agreed on a joint overall strategy for education monitoring through international, national, and state assessments as well as continuous reporting. Substantial funds are being invested in educational monitoring including a national academic institute set up by the federal states in addition to their individual state institutes. The BMBF has launched a comprehensive framework for empirical educational research that is currently in its second phase focusing on (1) increasing educational equality by identifying and developing individual potentials, (2) dealing with diversity and strengthening societal cohesion, (3) supporting quality in the educational system, and (4) designing and using technological developments in education (BMBF 2018). Both quantitative and qualitative research are being encouraged, and more emphasis is being given to the practical relevance of possible research findings for implementation in educational practice. Educational research has also been supported significantly in recent years by the strong increase in university chairs of empirical educational research. Despite some disputes between the

(traditionally more quantitatively and empirically focused) educational psychology and the (habitually more theoretically or qualitatively and empirically oriented) educational science, recent years have seen a productive dialogue and cooperation between the disciplines involved in education. This development ultimately led to the founding of an interdisciplinary Society for Empirical Education Research (GEBF) in 2012 as well as a new interdisciplinary open-access online journal (Journal for Educational Research Online, JERO) in 2009. Furthermore, evidence-based thinking has changed teacher training, instructional approaches, and school development in many ways.

Nevertheless, there have also been criticisms of the German educational monitoring strategy and developments over the last two decades. These include its focus on a few core domains with the potentially negative consequences for other subjects such as the arts and history in terms of, for example, appreciation, attention, effort invested in further development, or funding. Worries have also been expressed about schools and teachers using teaching-to-the-test strategies (Volante 2004). On a more general level, criticism has questioned the utility approach underlying the selection of domains and the resulting shift away from the idea of education for the primary sake of human development. This also raises the question of how to define the desired outcome of education and how performance skills such as reading and math relate to other outcomes of educational processes such as social, emotional, or personal skills; personality; attitudes; and motivational characteristics. Similarly, there are concerns that educational research itself has been mainstreamed into a service discipline for educational administration with research money and positions being awarded only to researchers and research closely linked to political interests and priorities (Bormann 2015). Regarding methodological issues, critics have expressed concern over the general ability of standardized tests to measure complex domains and the exclusion of entire sub-areas due to the lack of any opportunity to measure them in the current frameworks. Finally, yet importantly, the costs related to the international and national assessments have been criticized, arguing that these funds could otherwise be invested directly in the educational system and in improving its quality.

Educational monitoring and assessment currently face multiple challenges. One is the shift from paper-and-pencil to computer-based assessments. The new mode of assessment based on digital devices opens up new opportunities regarding item formats and which skill sets can be investigated. However, there are many methodological issues regarding trend analyses and construct (in)equivalence that have yet to be resolved. Looking at the assembly of LSA in Germany, it also becomes clear that despite the acknowledged importance of early education, the earliest assessment is performed in grade 4 (at age 10) and no (international) standardized assessments are being implemented in early childhood. Important new developments apart from the digitalization of assessments include increased interest in the implementation of school panels in LSA (see, e.g. the school panel within PISA 2000–2009 in Germany). Hence, for example, the school panels planned for PIRLS 2016 and 2021 will make it increasingly possible to combine evidence on school effectiveness with measures that are directly relevant for school development.

References

Autorengruppe Bildungsberichterstattung. (Eds.). (2018). *Bildung in Deutschland 2018. Ein indikatorengestützter Bericht mit einer Analyse zu Wirkungen und Erträgen von Bildung* [Education in Germany 2018. An indicator-based report with an analysis on the effects and earnings of education]. Bielefeld: wbv.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiß, M. (Eds.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* [PISA 2000: Basic competencies of students in an international comparison]. Opladen: Leske + Budrich.

Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The national educational panel study: Need, main features, and research potential. *Zeitschrift für Erziehungswissenschaft* [Journal for Education], *14*, 5–17.

Blossfeld, H.-P., Bos, W., Daniel, H.-P., Hannover, B., Köller, O., Lenzen, D., McElvany, N., Roßbach, H.-G., Seidel, T., Tippelt, R., & Wößmann, L. [Aktionsrat Bildung]. (2018). *Aktionsrat Bildung: Digitale Souveränität und Bildun*g [Action council education: Digital sovereignty and education]. Münster: Waxmann.

Bormann, I. (2015). *Unsicherheit und Vertrauen* [Uncertainty and trust]. *Paragrana*, *24*, 151–163.

Bos, W., Lankes, E.-A., Prenzel, M., Schwippert, K., Valtin, R., & Walther, G. (Eds.). (2003). *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich* [First results from PIRLS. Student achievements at the end of the fourth grade in an international comparison]. Münster: Waxmann.

Bos, W., Valtin, R., Hußmann, A., Wendt, H., & Goy, M. (2017). IGLU 2016: Wichtige Ergebnisse im Überblick [PIRLS 2016: Important results]. In A. Hußmann, H. Wendt, W. Bos, A. Bremerich-Vos, D. Kapser, E.-M. Lankes, N. McElvany, T. C. Stubbe & R. Valtin (Eds.), *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* [PIRLS 2016. Reading literacy of primary school students in Germany in an international comparison] (pp. 13–28). Münster: Waxmann.

Bundesministerium für Bildung und Forschung. (2018). *Rahmenprogramm empirische Bildungsforschung*. Bonn/Berlin: BMBF – Referat Bildungsforschung.

Hamre, B. K., & Pianta, R. (2010). Classroom environments and developmental processes: Conceptualization, measurement, & improvement. In J. L. Meece & J. S. Eccles (Eds.), *Handbook of research on schools, schooling and human development* (pp. 25–41). New York: Routledge.

Händle, C., Oesterreich, D., & Trommer, L. (Eds.). (1999). *Aufgaben politischer Bildung in der Sekundarstufe I. Studien aus dem Projekt Civic Education* [Tasks of political education at secondary level I. Studies from the project Civic Education]. Wiesbaden: Springer Fachmedien.

Hartwig, S., Schwabe, F., & McElvany, N. (submitted). *The interplay between teacher attitudes towards heterogeneity, differentiated instruction, and students' reading competence*.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Husén, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries, volume I–II*. Stockholm: Almqvist & Wiksell.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise* [The development of national educational standards. An expertise]. Bonn: Bundesministerium für Bildung und Forschung (BMBF).

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. J. (2004). *The development of national educational standards: An expertise*. Berlin: Federal Ministry of Education and Research (BMBF).

KMK – Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (1997). Grundsätzliche Überlegungen zu Leistungsvergleichen innerhalb der Bundesrepublik Deutschland – Konstanzer Beschluss – [Fundamental Considerations on Performance Comparisons within the Federal Republic of Germany – Constance Resolution –]. Retrieved online: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/1997/1997_10_24-Konstanzer-Beschluss.pdf [05.03.2020].

KMK – Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. [Comprehensive strategy by the Standing Conference for educational monitoring]. München: Luchterhand.

KMK – Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2016). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. [Comprehensive strategy by the Standing Conference for educational monitoring]. Köln: Wolters Kluwer.

Lang, M., & Schulz-Zander, R. (1994). Informationstechnische Bildung in allgemeinbildenden Schulen – Stand und Perspektiven [Information technology education in general education schools – status and perspectives]. In H.-G. Rolff, K.-O. Bauer, K. Klemm, H. Pfeiffer, R. Schulz-Zander (Eds.), *Jahrbuch der Schulentwicklung* [Yearbook of school development] (Vol. 8) (pp. 309–353). Weinheim: Juventa.

Lehmann, R., Peek, R., Pieper, I., & von Stritzky, R. (1995*). Leseverständnis und Lesegewohnheiten deutscher Schülerinnen und Schüler* [Reading comprehension and reading habits of German students]. Weinheim: Beltz.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Kennedy, A. M. (2003). *Trends in children's reading literacy achievement 1991–2001: IEA's repeat in nine countries of the 1991 Reading Literacy Study*. Chestnut Hill: Boston College.

McElvany, N., Kortenbruck, M., & Becker, M. (2008). Lesekompetenz und Lesemotivation: Entwicklung und Mediation des Zusammenhangs durch Leseverhalten [Reading literacy and reading motivation: Their development and the mediation of the relationship by reading behaviour]. *Zeitschrift für Pädagogische Psychologie* [Journal of Educational Psychology], *22*, 207–219.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 assessment framework* (2nd ed.). Retrieved from Boston College, TIMSS & PIRLS International Study Center website. http://timssandpirls.bc.edu/pirls2016/framework.html

OECD. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris: OECD Publishing. https://doi.org/10.1787/9789264195905-en.

Prenzel, M., Stadtler, M., Friedrich, A., Knickmeier, K., & Ostermeier, C. (2009). *Increasing the efficiency of mathematics and science instruction (SINUS) – A large scale teacher professional development programme in Germany*. Kiel: Leibniz-Institute for Science Education (IPN).

Schultze, W. (1975). *Die Leistungen im Englischunterricht in der Bundesrepublik im internationalen Vergleich* [Learning achievement in English in Germany in an international comparison]. Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung (Sonderheft Mitteilungen und Nachrichten).

Schultze, W., & Riemenschneider, L. (1967). *Eine vergleichende Studie über die Ergebnisse des Mathematikunterrichts in zwölf Ländern* [Learning achievement in math in Germany in an international comparison]. Frankfurt a.M: Deutsches Institut für Internationale Pädagogische Forschung (Sonderheft Mitteilungen und Nachrichten Nr. 46/47).

Stahns, R., Rieser, S., & Lankes, E.-M. (2017). Unterrichtsführung, Sozialklima und kognitive Aktivierung im Deutschunterricht in vierten Klassen [Classroom management, social climate, and cognitive activation in German lessons in fourth grade]. In A. Hußmann, H. Wendt, W. Bos, A. Bremerich-Vos, D. Kasper, E.-M. Lankes, N. McElvany, T. C. Stubbe & R. Valtin (Eds.), *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* [PIRLS 2016. Reading literacy of primary school students in Germany in an international comparison] (pp. 251–278). Münster: Waxmann.

Valtin, R. (2017). Einordnung der IGLU-2016-Befunde in das europäische Rahmenkonzept für gute Leseförderung. In A. Hußmann, H. Wendt, W. Bos, A. Bremerich-Vos, D. Kapser, E.-M. Lankes, N. McElvany, T. C. Stubbe & R. Valtin (Eds.), *IGLU 2016. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* [PIRLS 2016. Reading literacy of primary school students in Germany in an international comparison] (pp. 315–328). Münster: Waxmann.

Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy, 35*. Retrieved from: https://files.eric.ed.gov/fulltext/EJ848235.pdf

Wendt, H., Walzebug, A., Bos, W., Smith, D. S., & Bremerich-Vos, A. (2017). Germany. In I. V. S. Mullis, M. O. Martin, S. Goh, & C. Pendergast (Eds.), *PIRLS 2016 Encyclopedia: education policy and curriculum in reading*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/pirls2016/encyclopedia/

Chapter 13
The Hungarian Educational Assessment System

Ildikó Balázsi and László Ostorics

Since 1968, when it joined the International Association for the Evaluation of Educational Achievement (IEA), Hungary has participated in approximately 25 international large-scale student assessments. Participation in these assessments and the development of a national assessment system are intended to inform educational policy makers, professionals and the public. This chapter presents the history and the current state of the Hungarian assessment system with special focus on international studies and the National Assessment of Basic Competencies as its main pillars. The chapter's main focus is on equity, a key issue in Hungarian public education, which our assessment data sheds light on. Results from PISA 2015, TIMSS 2015, PIRLS 2016 and NABC 2017 are used to explore differences between schools in terms of socio-economic status and academic achievement, as well as the strength of relationship between the former and the latter. Policy recommendations regarding assessment and the public education system as a whole are made.

Introduction to the International Assessment Context and Its History in Hungary

In Hungary, compulsory schooling (including 3 years in kindergarten) lasts from age 3 to age 16 in accordance with the Act on National Public Education (2011), and the public education system is based on 8 years in primary school and 4–5-years in

I. Balázsi (✉)
Department of Public Education Analysis, Hungarian Educational Authority, Budapest, Hungary
e-mail: balazsi.ildiko@oh.gov.hu

L. Ostorics
Department of Assessment and Evaluation, Hungarian Educational Authority, Budapest, Hungary

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_13

secondary school. Students, however, may change to grammar schools as early as after 4th grade – the system offers grammar schools of 8, 6 and 4 grades, of which the last-named is the most popular. All academic-track secondary schools end with the Matura exit exam at the end of 12th grade that works as a university admissions exam as well. Vocational training can be started after 8th grade either in vocational secondary schools that offer also Matura or in vocational schools that focus on occupation-related qualifications.

According to the Hungarian Central Statistical Office in 2011, at the time of the last population census in Hungary, approximately 68% of the young adults aged 20–24 finished their secondary education with a Matura examination, 16% had vocational qualification without a Matura and 16% had not finished upper secondary education (Hungarian Central Statistical Office n.d., Table 2.1.1). The proportion of early school leavers, the share of the population aged 18–24 with at most lower secondary education who were not involved in any education or training during the 4 weeks preceding the survey, was 12.5% in 2018 according to Eurostat (Eurostat n.d., Table SDG_04_10).

International and National Studies Before the Millennium

Hungary joined the International Association for the Evaluation of Educational Achievement (IEA) in 1968, shortly after IEA had successfully conducted its first international large-scale student assessments, the Pilot Study and the First International Mathematics Survey. Hungary was the first Eastern European communist country to join the IEA and in fact the only one from the Soviet bloc up to its collapse in 1990 (Brassói and Kádár-Fülöp 2011); although Poland and Yugoslavia did participate in the first IEA Pilot Study in 1960, and Poland and Romania sought to join some parts of the Six Subject Survey in 1970–1971. Unfortunately, Poland did not manage to finish the survey with a published dataset and Romania only participated in the French as a foreign language part of the study. Only Hungary had permanent and almost full participation in the tests.

Hungary took part in a couple of IEA studies before the millennium, including the Second International Science Study, the Second IEA Study on Reading Literacy and TIMSS. Hungarian researchers and policy makers considered mathematics, science and reading comprehension as essential domains of interest from early on (Kádár-Fülöp 2015). Additionally, Hungary joined many other innovative areas of research conducted by IEA, for example, studies about civic education, ICT skills and composition skills (for a detailed list of IEA studies in which Hungary participated, see Brassói and Kádár-Fülöp 2011, Table 1, p. 436).

The membership in IEA had a stimulating effect on the Hungarian educational research community (Brassói and Kádár-Fülöp 2011). Báthory (1992) identified three areas of benefits for the educational system: (1) methodological advances, (2) the ‘window effect’ and (3) introduction to system-level analysis.

Peer-learning on fields such as conceptualization of educational assessments, framework development, field operations, data processing and analysis resulted in methodological advances in developing a national assessment system (Halász and Lukács 1987; Kádár-Fülöp 2015). Besides the expertise coming from working day-to-day on the national implementation of the studies, IEA organized workshops and trainings from its early days, and Hungarian educational researchers benefited a lot professionally from these occasions too. As an example, articles about the subsequent Hungarian reforms on the national curriculum usually mention the International Curriculum Seminar held in Gränna in 1971 as having a long-lasting effect on researchers involved in those reforms (Ballér 2001; Brassói and Kádár-Fülöp 2011). Hungarian experts were also able to join in the work of IEA and contribute to the development of the international assessments. Tamás Varga, the leading figure behind the renewal of the Hungarian mathematics curriculum and the 'new math' movement was a member of the International Mathematics Committee of the Second International Mathematics Study (Travers 2011). Zoltán Báthory, the representative of Hungary in IEA's Standing Committee until 1994 and an honorary member of IEA, contributed to the mathematics and science assessments of IEA in various ways.

The 'window effect', meaning the possibility of international research cooperation in an era of strict ideological isolation, had its merits alongside the direct professional development of researchers. Through IEA studies and regular meetings with researchers all over the world, Hungary had the possibility of viewing its own educational system in a wider, global context (Báthory 1992; Kádár-Fülöp 2015).

The third and the most evident effect of IEA studies was the introduction of system-level analyses in Hungary. Although the IEA studies were unprecedented sources of comparable data on many aspects of educational systems worldwide and IEA always emphasized that the studies were meant to research the processes and methodological differences leading to different student outcomes, the main interest of the Hungarian politicians and educational professionals alike was the overall achievement of the system (Kádár-Fülöp 2015). The international and national reports of IEA studies up to 1990 showed good results in mathematics and science, and poor results in reading comprehension. The latter shocked the Ministry of Education and experts as well. Methodological changes and the liberalization of methods of teaching reading from the late 1970s were linked directly to these IEA findings (Báthory 1992). The liberalization processes operated in parallel with other endeavours towards freedom of education (e.g. free school choice for parents and students, more autonomy for schools in relation to the selection of methods and materials, the right for non-governmental organizations to maintain primary and secondary schools) and resulted in new regulations such as the 1985 Public Education Act ("Az oktatásról. 1985. évi I. törvény" 1985).

Researchers and institutes working on IEA studies were also involved in developing a national assessment system (Halász and Lukács 1987; Kádár-Fülöp 2015). The first nationally representative sample-based study, TOF-80 assessed 4th and 8th grade students' abilities in various subject areas (Báthory 1983). Though it was a

stand-alone study with no follow-up, the experiences benefited the planning and implementation of the Monitor Studies. The Monitor Studies, first implemented in 1986, were created with the goal to regularly monitor student achievement in reading, mathematics, science, information technology and cognitive skills. The study was first repeated in 1991 and from that time it became a bi-annual study until its termination in 2005. The Monitor Studies complemented IEA studies with trend data on student achievement. Whenever it was possible, the Monitor Studies used the same samples as IEA studies (Vári 1997; Vári et al. 2000).

International and National Assessments in the New Millennium

With the long history of international student assessments, it was no surprise when Hungary joined the OECD PISA project when it was launched in 1997. PISA 2000 results were received with a sense of disappointment in Hungary, because not only reading literacy, but also mathematics and science literacy performance were mediocre compared to previous good results in IEA studies (Kádár-Fülöp 2015). The seemingly contradictory results of IEA's PIRLS and TIMSS and OECD's PISA were and are in the focus of much attention in Hungary since then. The first article about Hungarian PISA 2000 results gave various possible explanations for the differences, for example, the differences in standardization and populations (the computation of average scores in IEA studies vs. the OECD average in PISA, grade-based vs. age-based sample, 4th and 8th graders vs. 15-year-olds); different emphases of the studies (research on educational processes vs. indicators of the quality of education); differences in the frameworks, especially the contexts in which the scientific or mathematical problems are embedded (textbook-like stems in TIMSS items, based on curricula vs. situations from everyday life in which students have to use their mathematical and scientific abilities to solve the problems in PISA) (Vári et al. 2002). The OECD's more intense communication and 'advertising' of PISA and the poorer results of Hungary in PISA both contributed to the greater attention among the media and the general public following PISA and other international student assessments in the new millennium (Kádár-Fülöp 2015).

Based on experiences from the international assessments and the Monitor Studies, a new, annual student assessment system was initiated in Hungary in 2001 (Berényi et al. 2013). The National Assessment of Basic Competencies (NABC) is a constantly developing system, and during the more than 15 years of its existence the characteristics of the study have changed considerably. However, the basic aims of the study did not change a lot from the beginning; the main aim was and is to give schools objective, nationally comparable data on their students' abilities in important literacy domains (Ostorics 2015). Based on that, every year all students in grades 6, 8 and 10 participate in these assessments of reading and mathematics competencies. School Reports have been available to schools – and from 2007 on to the general public as well – in order to help them evaluate their results. Results are

reported not only in absolute terms, but also in comparison to students' socio-economic status (from 2003) and to earlier results (from 2010) as well (Balázsi 2016).

Despite the evident effects of IEA studies mentioned before, researchers felt that the full potential of the studies has never been reached in Hungary. For example, Brassói and Kádár-Fülöp (2011) wrote that "In spite of our zealous participation in several IEA surveys, their results had little direct impact on education policy or instruction in Hungary" (pp. 435–436). The several merits of participation in international assessments have been overshadowed by the fact that during our long history, policy changes and system development initiatives were rarely able to improve the results or equity of the system.

International Assessments Today

Today, international assessments are a constant element of the Hungarian assessment system and provide a wider perspective alongside periodic national assessments and exams of various goals, stakes and scope that cover the span of primary and secondary schooling (Fig. 13.1.).

National assessments and exams range from no-stake university-developed diagnostic programmes to state-run centrally developed high-stakes examinations. Diagnostic assessments are available to schools from grade 1 to grade 6. These are developed by Szeged University's Centre for Research on Learning and Instruction and cover domains such as prerequisite knowledge fields of school readiness for first graders delivered via paper-based tests (writing-movement coordination, relational

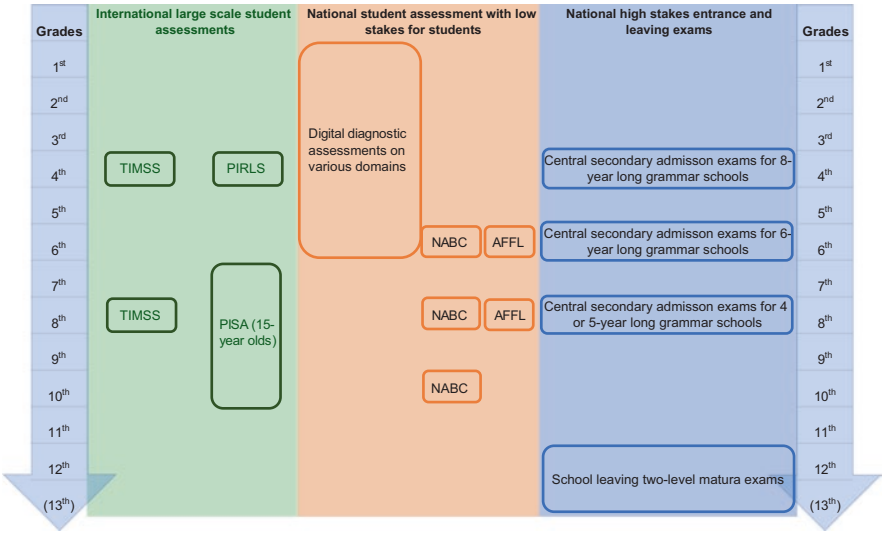

Fig. 13.1 The national assessment system in Hungary

vocabulary, basic calculations, experimental deduction, basic social skills) and digital reading, mathematics and science test tools for higher grades. Schools are legally required to utilize the former in cases when kindergarten reports or the experiences of the first month of schooling compel them to do so (Centre for Research on Learning and Instruction n.d.).

Centrally developed and legally full-cohort albeit low-stake assessments are the National Assessment of Basic Competences (NABC) and the Assessment of First Foreign Language (AFFL), both the responsibility of the Department of Assessment and Evaluation in the Educational Authority (Szabó et al. 2018). These are administered as pen-and-paper tests. The NABC assesses non-curriculum-based domains including reading and mathematics literacy and has a well-developed and extensive on-line administration and reporting system. Reports are available from student to national level and provide mean scores and distributions of students across proficiency levels per grade, school type and cognitive domain compared to the national average and to the results of different subpopulations (Ostorics 2015). Besides, the reports communicate value-added analyses outcomes as students' results are compared to their expected results based on their socio-economic status and on their earlier results where available. Software is also provided to schools to allow for further analysis of their own results. Compared to the NABC, the AFFL is a limited instrument. Students in 6th and 8th grades are legally required to take the curriculum-based test; however, central administration and reports are still lacking.

Unlike national tests, participation in international assessments is not obligatory for schools, students or teachers. Still, participation rates are constantly high in TIMSS, PISA and PIRLS, the three major international assessments that Hungary participates in since their first administration. Although not an achievement study, Hungary took part in OECD TALIS 2018 involving teachers in ISCED 2 (lower secondary) schools as well.

As seen, assessments examine the educational system at various points. One way of looking at these points is taking note of the grades and school types involved. At the stage when PIRLS is administered, at grade 4, all students attend primary schools. In fact, this is the last point in Hungarian schooling when all students go to a single type of school. The 6th grade NABC assesses students in lower secondary and in 8 grade grammar schools. TIMSS and NABC in the 8th grade look at students in lower secondary and 6 and 8 grade grammar schools, while PISA assesses 15-year-old students from 7th to 10th grade in lower secondary schools (grade 7 and 8), all three tracks of academic grammar schools, and the two tracks of vocational training. The 10th grade NABC assesses students is all secondary tracks.

A set of international, state and local assessments administered in the same national educational context does not necessarily constitute a system. In the case of Hungary, however, the implementation of the student measurement identifier (SMID) provides a possibility to link the datasets yielded by such diverse sources. The SMID was implemented in the 2007/2008 academic year and has the dual goal of protecting student privacy and allowing for the examination and analysis of relations between results of national tests (such as the NABC) and international tests (such as TIMSS, PISA and PIRLS) (Ostorics 2015).

An Example: The Issue of Equity in the Hungarian School System

International assessment results are usually the focus of attention because of the rankings and national mean scores. As noted above, IEA and OECD programmes provide seemingly contradictory results in Hungary: PIRLS and TIMSS outcomes are always above international average and are sometimes excellent, while the results of Hungarian 15-year-olds rarely attain the OECD average in PISA. The latter showed further decline with the shift from printed to digital assessment mode in 2015. While rankings are attracting more attention, the main lessons learnt from international assessments are related to equity.

Issues of equity were in the focus from the first IEA assessments in Hungary. Báthory (1992) cites a table presented at the 1986 IEA General Assembly meeting by Sixten Marklund showing how the variation in the science performances of students is distributed between school and student levels in different countries in 1970. Among the nine countries presented there, Hungary had the highest between-school variance portion (40%, the country average was 26%) in grade 4 and above-average between-school variance (34%, the country average was 29%) in grade 8. Also, the effect of the socio-economic status on student performance was around the IEA average in reading and somewhat below the IEA average in science, with several countries having a lower impact of SES on performance (Báthory 1992). These findings were particularly disturbing for Hungary, where the declared socialist state policy was egalitarian, and equality of opportunity – the educational and social mobility of working class children – was one of the most important aims of the educational system.

Almost 50 years later, after fundamental changes in the political and educational systems, equity is still a serious problem in Hungary. PISA and NABC have shown that differences between schools in terms of the socio-economic status and the performance of students are still considerable. About 25–36% of the differences in student performance in primary and lower secondary level come from differences between schools (Table 13.1). For the upper secondary level, the differences between schools are even more pronounced; above 50% of the variance originates from school-level differences in all domains of PISA and NABC. These values are high compared to other countries, with the PISA estimates for Hungary well above the corresponding OECD averages of between-school variance proportions (Organisation for Economic Cooperation and Development 2016b, Table I.6.9, p. 409). The increase in between-school variances between grades 8 and 10 originates from the structure of the educational system as described earlier, including the selection of students for academic secondary schools.

There are considerable differences in students' socio-economic status between schools as well. PISA's index of social inclusion (the proportion of variance coming from differences within schools) is one of the lowest for Hungary at 62.6% (OECD average is 76.5%), meaning that the intra-class correlation (rho), a measure of between-school variance in SES, is one of the highest (Organisation for Economic

Table 13.1 The differences between schools and the effect of socio-economic status on the performance in various studies

Student population	Source of data	The percent of variance coming from differences between schools				The strength of the relationship between the SES and the performance (percentage of variance in student performance explained by the socio-economic status, R^2)		
		SES[a]	Reading	Mathematics	Science	Reading	Mathematics	Science
Grade 4	PIRLS 2016	41.3	27.2			31 (2.1)		
	TIMSS 2015	31.0		24.8	28.0		33 (2.0)	32 (2.0)
Grade 6	NABC 2017	45.1	27.6	29.8		28 (0.3)	24 (0.3)	
Grade 8	TIMSS 2015	32.1		36.1	32.9		37 (2.1)	32 (2.1)
	NABC 2017	45.5	29.8	32.9		28 (0.3)	27 (0.3)	
15-year-olds	PISA 2015	37.4	58.4	53.7	55.4	22 (1.5)	21 (1.5)	21 (1.4)
Grade 10	NABC 2017	50.1	52.8	51.5		28 (0.3)	26 (0.3)	

Sources: Organisation for Economic Cooperation and Development (2016a), International Association for the Evaluation of Educational Achievement (2016, 2017), and Oktatási Hivatal (2018)

[a]Socio-economic status is measured similarly, but not exactly with the same variables and methods in the different studies. For the above computations, we were using the index provided by the studies for measuring student SES. In PIRLS and TIMSS grade 4 the Home Resources for Learning index was used, in TIMSS grade 8 the Home Educational Resources index, in NABC the Family Background index, in PISA the Economic, Social and Cultural Status. For a detailed description of the indices, see the sources above

Standard errors of estimates are shown in parentheses

Between-school variance proportions are equal to the intra-class correlations of the empty two-level model (students in schools) for PISA and the intra-class correlations of the empty three-level model (students within classes within schools) for TIMSS, PIRLS and NABC

Cooperation and Development 2016b, Table I.6.10, p. 410). Hence, social segregation is high in the Hungarian education system with schools differing considerably on their student intake and having more homogenous student populations within schools. Moreover, the index of social inclusion is already low on the primary level. Based on the Family Background index of NABC 2017, between 45% and 50% of the variance of students' socio-economic status comes from differences between schools in all grades. According to this, while tracking in Hungary on the secondary level increases the academic segregation, social segregation is already large in lower grades and does not increase much more when further tracking takes place.

It should be noted that in Hungary parents are free to choose the primary school for their children since the 1985 Act I on Public Education. While this may further

school competition, which helps with matching school offers to student demands, it might be equally detrimental for equity (Musset 2012). Free school choice favours families who have sufficient resources to take into account criteria other than low expenses and short distance from home when choosing a suitable school for their children (Organisation for Economic Cooperation and Development 2014).

The relationship between students' socioeconomic status (SES) and performance is also high in Hungary and does not change considerably between grades (Table 13.1). In PISA, the effect of the ESCS (index of the economic, social and cultural status) is above the OECD average, and the strength of the relationship between student ESCS and performance is one of the strongest among PISA 2015 countries (Organisation for Economic Cooperation and Development 2016b, Table I.6.12a, p. 412). Given the high levels of segregation in the Hungarian educational system, it is not surprising that the relationship between the economic, social and cultural status and performance of students is most evident at the school level. While the effect of students' ESCS within schools is small, 80.1% of variation between schools in science performance is explained by students' and schools' ESCS in Hungary. Overall, taken together, student and school ESCS explain 43.4% of the variation in student performance, which is the highest value in PISA 2015 and almost twice the OECD average 22.4%.

Because of the high social disparities in the performance of students and schools in Hungary and the high proportions of low achievers, research on resiliency (the ability of students to succeed in school against the adversities that may arise from coming from a disadvantaged family background) is very important for us. International and national student assessments are excellent sources of data for that purpose as well. PISA has an interest in student resiliency since the 2006 cycle. In research on factors associated with resilience, or the ability of disadvantaged students to perform well, the OECD found that in Hungary and in many other countries, students' self-efficacy and the number of hours students reported spending on regular lessons at school learning science were related to student resilience in science (Organisation for Economic Cooperation and Development 2011). Other factors, such as general interest in science, participation in science-related activities, competitiveness and selectivity of the schools based on academic record or the activities of schools to promote the learning of science and quality of school educational resources were not significantly related to resiliency.

Agasisti et al. (2018) reported that, after controlling for student gender, the ESCS of students and schools, and differences between the language spoken at home and the language of instruction, schools with a more positive school climate and lower absenteeism were significantly better in promoting resiliency in PISA 2015 in Hungary. In contrast, school resources indicators were not connected to resiliency. The effects of the ratio of computers to students and the average class size were not statistically significant, while the number of extracurricular activities at school had a negative relationship with resiliency. The last two had a positive effect in many other countries and on average among OECD countries too, which highlights the need to examine school-level factors in a national context.

The reports on the NABC strive to take differences in student intake between schools into account and to show how schools are able to deal with student groups with various social and cultural backgrounds. As a first step to examine school level factors fostering resiliency, the Educational Authority identifies schools that score significantly above what is estimated based on their students' socio-economic status or previous achievement. A non-ranking list of these schools has been published annually since the 2014–2015 school year (Educational Authority 2018).

Policy Recommendations

The Hungarian national assessment system currently focuses on the domains of reading literacy, mathematics and science. Among international assessments, only PISA has an option to assess other domains. Participation in other international large-scale student assessments (e.g. in IEA studies of civic and citizenship education or computer and information literacy) might help to plan curriculum development and reforms of teaching and learning materials, practices, etc. in those specific fields.

However, current evidence based on ongoing international and national assessments is plentiful concerning the structural, quality and equity issues of the Hungarian educational system, so limited resources are probably better used in disseminating outcomes, engaging in secondary analysis of existing databases and making policy recommendations based on evidence coming from them. The student measurement identifier (SMID) provides the possibility of linking the data of international assessments to the NABC and other national data sources at the student level. This, in turn, expands the research and policy topics that could be addressed, for example, by following up students' educational career from grade 4 to the end of their tertiary education.

PISA, TIMSS and PIRLS are pillars of the Hungarian assessment system and continuing participation in them provides invaluable trend data on the quality and characteristics of the educational system with the possibility of international comparison and opportunity to learn from good practices from all over the world. The long history of participating in international large-scale studies and the constant development of Hungary's national monitoring and assessment system yield data that can inform policy makers on the structural changes that may be beneficial for Hungarian public education, allow for the identification of regional, local and school-level issues, and offer educational researchers the possibility to conduct in-depth analysis. The main lessons are summarized below.

Achievement indicators provided by international assessments show that Hungarian students at the beginning of lower secondary education are more proficient than their peers in many other education systems, while 15-year-olds' results lag behind the OECD average in all assessment domains, even more so when the data are collected using digital tools. This raises the question of the quality of lower

secondary schooling, how we can better support literacy development of children, and which policy initiatives can improve teaching and learning in lower secondary schools.

Free school choice seems to affect the student intake of the public education system in a way that results in schools which are homogenous in not just academic achievement but in socio-economic status as well. This hinders equity, favours well-off families and does not facilitate the prevention of early school leaving. Policies to mitigate the adverse effects of the current completely free school choice regulations should be explored.

The system of national assessments needs to adapt to serve demands that include supporting early drop-out prevention programmes. Currently, several projects financed by EU structural funds or the state of Hungary aim to reach the above-stated goals. These projects are overseen by the Educational Authority, as the state agency responsible for collecting and reporting assessment data.

As for preventing early dropout and fostering resilience, it should be noted again that schools that can constitute a basis for these processes can be identified by the means of the NABC reports and databases. A considerable number of schools can achieve significantly better results than that estimated based on their students' ESCS-index or their students' previous proficiency scores. The existence of such schools indicates that the effects of socio-cultural and socio-economic disadvantage can be mitigated, even in the current structure of the education system. A relevant further step is examining school-level factors and exploring the good practices that may result in student resiliency. On the other hand, schools that do not attain expected levels based on their students' socio-cultural status or previous results also can be identified with the use of NABC data. One of the main goals of the projects mentioned above is to build tools to facilitate communication between schools to promote networked learning.

Other relevant goals include promoting the use of digital technologies and moving the national assessment system online, adding new domains to the existing ones, including science and digital literacy, and fostering the use of existing data by making it more accessible to both the general public and professionals.

References

2011évi CXC. törvény a nemzeti köznevelésről [Act CXC of 2011on National Public Education]. (2011). Retrieved from http://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=A1100190.TV

Agasisti, T., Avvisati, F., Borgonovi, F., & Longobardi, S. (2018). *Academic resilience: What schools and countries do to help disadvantaged students succeed in PISA* (*OECD education working papers, 167*). Paris: OECD Publishing. https://doi.org/10.1787/e22490ac-en.

Az oktatásról. 1985. évi I. törvény. (1985). *Magyar Közlöny, 19. sz.*, pp. 461–492.

Balázsi, I. (2016). Applying value-added models to student achievement. *Magyar Pedagógia, 116*(1), 3–23.

Ballér, E. (2001). Új tendenciák a tantervelméletben és a tantervfejlesztésben. *Iskolakultúra, 2001*(9), 67–72.

Báthory, Z. (1983). Az iskolai nevelés néhány összetevőjének vizsgálata egy felmérés tükrében (TOF-80) – Bevezető. *Pedagógiai Szemle, 33*(2), 135–139.

Báthory, Z. (1992). Hungarian experiences in international student achievement surveys. *Prospects, 22*(4), 434–440.

Berényi, E., Bajomi, I., & Neumann, E. (2013). Une expérience hongroise visant à gouverner par les nombres – genèse et évolution du Système National de Mesure des Compétences. In C. Maroy (Ed.), *L'école à l'épreuve de la performance* (pp. 71–88). Brussels: De Boeck Supérieur.

Brassói, S., & Kádár-Fülöp, J. (2011). How IEA influenced the education system in Hungary. In C. Papanastasiou, T. Plomp, & E. C. Papanastasiou (Eds.), *IEA 1958–2008: 50 years of experiences and memories* (Vol. 1, pp. 431–446). Amsterdam: The International Association for the Evaluation of Educational Achievement (IEA).

Center for Research on Learning and Instruction, University of Szeged. (n.d.). *Developing diagnostic assessments*. Retrieved from http://edia.hu/projekt/?q=en

Educational Authority. (2018). *Kiemelkedő teljesítményű iskolák.* Retrieved from https://www.oktatas.hu/kozneveles/meresek/kompetenciameres/kiemelkedo_teljesitmenyu_iskolak

Eurostat. (n.d.) *Eurostat Database. Education and training outcomes*. Retrieved from https://ec.europa.eu/eurostat/web/education-and-training/data/main-tables

Halász, G., & Lukács, P. (1987). *Az IEA magyarországi hatása: oktatáspolitikai esettanulmány* (*OPI Értékelési Központ Közleményei, 14*). Budapest: Országos Pedagógiai Intézet.

Hivatal, O. (2018). *OKM 2017 Kutatói Adatbázis*. Budapest: Oktatási Hivatal.

Hungarian Central Statistical Office. (n.d.) *Population census 2011, educational data*. Retreived from http://www.ksh.hu/nepszamlalas/tablak_iskolazottsag

International Association for the Evaluation of Educational Achievement. (2016). *TIMSS 2015 international database.* Retrieved from https://timssandpirls.bc.edu/timss2015/international-database/

International Association for the Evaluation of Educational Achievement. (2017). *PIRLS 2016 international database.* Retrieved from https://timssandpirls.bc.edu/pirls2016/international-database/index.html

Kádár-Fülöp, J. (2015). Nemzetközi tudásszintmérés – hazai oktatáspolitika. *Education, 24*(2), 9–15.

Musset, P. (2012). *School choice and equity: Current policies in OECD countries and a literature review* (*OECD education working papers*, *66*). Paris: OECD Publishing. https://doi.org/10.1787/5k9fq23507vc-en.

Organisation for Economic Cooperation and Development. (2011). *Against the odds: Disadvantaged students who succeed in school*. Paris: OECD Publishing.

Organisation for Economic Cooperation and Development. (2014). *When is competition between schools beneficial?* (*PISA in focus, 42*). Paris: OECD Publishing. https://doi.org/10.1787/5jz0v4zzbcmv-en.

Organisation for Economic Cooperation and Development. (2016a). *PISA 2015 international database.* Retrieved from http://www.oecd.org/pisa/data/2015database/

Organisation for Economic Cooperation and Development. (2016b). *PISA 2015 results (volume 1). Excellence and equity in education*. Paris: OECD Publishing. https://doi.org/10.1787/9789264266490-en.

Ostorics, L. (2015). A tanulói teljesítménymérések jellemzői, jövőbeni irányvonalai, kritikái. In K. Széll (Ed.), *Mit mér a műszer?* (pp. 37–62). Budapest: Oktatáskutató és Fejlesztő Intézet.

Szabó, L. D., Szepesi, I., Takács-Kárász, J., & Vadász, C. S. (2018). *Országos kompetenciamérés 2017. Országos jelentés*. Budapest: Oktatási Hivatal. Retrieved from https://www.oktatas.hu/pub_bin/dload/kozoktatas/meresek/orszmer2018/Orszagos_jelentes_2017.pdf

Travers, K. J. (2011). The Second International Mathematics Study (SIMS): Intention, implementation, attainment. In C. Papanastasiou, T. Plomp, & E. C. Papanastasiou (Eds.), *IEA 1958–2008: 50 years of experiences and memories* (Vol. 1, pp. 431–446). Amsterdam: The International Association for the Evaluation of Educational Achievement (IEA).

Vári, P. (Ed.). (1997). *Monitor '95. National assessment of student achievement*. Budapest: Országos Közoktatási Intézet.

Vári, P., Bánfi, I., Felvégi, E., Krolopp, J., Rózsa, C., & Szalay, B. (2000). A tanulók tudásának változása I. A Monitor '99 felmérés előzetes eredményei. *Új Pedagógiai Szemle, 50*(6), 25–35.

Vári, P., Aux-Bánfi, I., Felvégi, E., Rózsa, C., & Szalay, B. (2002). Gyorsjelentés a PISA vizsgálatról. *Új Pedagógiai Szemle, 52*(1), 38–65.

Chapter 14
Educational Assessment in Iceland

Meyvant Þórólfsson

Introduction

Iceland is a thinly populated island in the North Atlantic Ocean close to the Arctic Circle. The population is about 1/3 of a million dispersed along its coasts where the south-west part is most densely populated with the city of Reykjavik at its centre. Covering an area of 103,100 km^2 , the island is geologically located on the Mid-Atlantic Ridge that separates the Eurasian plate and the North American plate. Thus, Iceland is among the most active volcanic areas of the world. Glaciers cover one-tenth of the island along with lots of rivers, waterfalls, geysers and fiords. The climate is warmer than the northern latitude indicates because of the warm Gulf Stream, and the bright summer nights; nevertheless, the climate is described as windy, cloudy and unstable.

For ages Iceland was a poor Danish colony, but in 1918 it received its first recognition as an independent state and in 1944 full independence was announced and The Republic of Iceland was founded. It developed rapidly due to beneficial factors, for example, the entry into the so-called Marshall plan after the Second World War, and consequently important economic, technological and scientific advances connected to fishing industries and eventually technology and innovation in other industrial sectors.

The legacy of literature has occupied the lives of Icelanders since long before Gutenberg introduced the printing press in the 1400s. Due to a variety of circumstances, nearly all the ancient literature was written in the vernacular language although most scholars and many intellectual farmers knew Latin by heart. For centuries, literacy has been considered high in Iceland and according to researchers such as Gíslason (1977) and Proppé (1983), specific circumstances have sustained the literary tradition, namely the climate and the northern location of the island with

M. Þórólfsson (✉)
University of Iceland, Reykjavik, Iceland
e-mail: meyvant@hi.is

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_14

its long and dark winter periods. People had few other choices than staying indoors where families assembled during long winter nights on evening wakes (i. kvöldvökur). Such wakes were a significant cultural tradition dating since the middle ages and lasting until the first half of the twentieth century. They involved various kinds of intellectual activities such as loud reading of the ancient sagas, poetry, rhymes, Bible reading and telling ghost stories.

During the middle ages, monasteries provided schooling and some priests and farmers also had schools in their homes (Guttormsson 1981; Proppé 1983). Cathedral schools became grammar schools after the Protestant Reformation in 1550, though not intended for the public but an elite preparing for priesthood or judicial practice. Public education did not receive much attention until the eighteenth century when the first law addressing public education came into act in 1880 and in 1907 the state agreed on providing public schools for children aged 10–14 years. By 1900, the first secondary public schools were also founded. Eventually, the basic policy was confirmed that all children should have equal opportunities to acquire basic education without any discrimination. At the compulsory level (elementary and lower-secondary, ages 5–16), education has been totally free for more than a century and now pupils are provided with all learning materials and resources.

Since the establishment of public schooling, assessing pupils and their learning has been organised and carried out by schools as part of the implemented curriculum and concurrently by educational authorities as part of the intended curriculum. Before 1970, the words evaluation (i. mat) and assessment (i. námsmat) were not found in educational discourse in Iceland. Instead, educators talked about tests (i. próf) and grades (i. einkunnir). Now the word assessment (i. námsmat) is almost exclusively used. Furthermore, discourse about different purposes of assessment emerges increasingly, i.e. assessment of learning, assessment for learning and assessment as learning.

Assessment, Testing and Grading

Influenced by the spirit of pietism during the eighteenth century and first half of the nineteenth century, priests were responsible for education and judging pupils' learning. Pupils received marks based on how well they had learned their lessons (Guttormsson 2008). After 1860, schools began to be established and gradually teachers became responsible for public education together with priests. Each day, marks were written in protocols and kept as data about learning and learning processes (Proppé 1983).

As the twentieth century commenced, most schools changed their daily grading to weekly grading, but most importantly the grades were still based upon subjective judgements of teachers and priests (Proppé 1983). The first signs of summative assessment appeared through a debate about yearly spring examinations. According to the Law on the Education of Children from 1907, pupils were to be tested 'orally' each spring in the traditional subjects, such as reading, religion, arithmetic, history,

zoology, crafts and physical education. According to regulations based on the 1907 law, grading was supposed to embody a number scale from 1 (bad) to 8 (excellent) (Proppé 1983).

By the end of the second decade of the twentieth century, ideas of psychometric methods and written tests began to emerge. In 1920, a distinguished Icelandic scholar, Steingrímur Arason, came home from his studies at Columbia University, New York, influenced by Edward Lee Thorndike's educational psychology. Arason's introduction of quantitative measurements and written tests induced immense controversy, which was no surprise because the battle between progressive thinkers and traditionalists was striking at that time. One of the largest compulsory schools in the country, Austurbæjarskóli, publicly declared itself as a progressivist school. Its principal, Sigurður Torlacius, had received his education in Europe and became familiar with 'The New School' and thinkers like Maria Montessori in Italy, Ovide Decroly in Brussels, John Dewey in America and members of the 'active school' (g. Arbeitschule) in Germany. Torlacius maintained that testing and grading practices were misleading because they focused on trivial skills instead of other important competences. He expressed his school's policy about testing and grading this way:

> Instead of the spring examinations we need process evaluation by school specialists … Instead of the grades, we should mainly show what the children have done … besides that, we should have personal communications between the teacher and the home, through which information about the children can be given both ways. (Thorlacius 1932, p. 23)

But his suggestions did not receive much support, so as Ellen C. Lagemann (1989) put it, 'Thorndike won and Dewey lost'. As the British philosopher of education R. S. Peters identified (as cited in Walker and Soltis 2009, p. 14), such high-sounding aims were commitments to certain values, but their role in everyday activities of teachers turned out to be insignificant.

Steingrímur Arason managed to convince most eminent scholars that the 'new testing methods' would secure reliable judgements about learning and one of them added that most importantly they would secure fairness and equity (cf. Hjörvar 1921). Still, there were those who feared psychometric tests and conceived them as dogmatic and that trying to measure 'cultural and social dimensions with quantitative measurements' out of context was unwise (Proppé 1983, p. 267). But ultimately the general agreement was that educational authorities should provide centralised written examinations because the old methods were considered too subjective and useless for comparison. Arason himself argued that it was time to provide opportunities for comparison between and within schools. The new methods came into use in most schools in the 1920s and in 1929 the first national tests were introduced. In the coming years and decades, centralised testing, though not standardised, earned its place as the mainstream way of assessing learning.

In 1946, the first law for one unified school system in Iceland was passed, *The Education Act of 1946*. Included were centralised examinations compulsory for grades 4, 6 and 8 and centralised entrance examination for grammar schools after grade 9, the National Examination (i. Landspróf). It was optional and at first very few students of each cohort passed it, 7% in 1950, 17% in 1965 and 25% in 1975.

Gradually, scholars began to worry about the negative influences that the 'Landspróf' had on the whole school system. Though it was originally meant as egalitarian means to secure equal rights for everyone, it gradually involved constricting effects with its emphasis on mere knowledge in traditional subjects. The focus was solely on book learning in subjects such as Icelandic, English, Geography, Mathematics and Physics. An eminent school administrator and educational advisor argued that schools should normally be organised bottom-up. But he asked if the academic emphasis and influence of the 'landspróf' had turned things upside down: 'Are the schools not shaped top-down instead? Do learning conditions and organisation of secondary schools not indeed control what is done in primary and lower-secondary schools?' (Gunnarsson 1963)

According to a new Act on the Comprensive Primary School that came into action in 1974, the assessment discourse finally took new directions. As the following paragraph indicates, the discourse about assessment was gaining a different momentum:

> Assessment of learning should not only be practiced at the end of a learning unit, rather it should be among the continuous activities of the school practice, entirely integrated with learning and teaching. The main purpose of assessment of learning is the motivation of students and learning assistance. (Law on the Comprensive Primary School 1974)

And a pamphlet from the Ministry denoted:

> Assessment has received increasing attention worldwide. At the same time focus on the nature and needs of individual students has increased and the learning process receives no less attention than the product of learning. (ME 1979, p. 3)

For two decades from 1970 to 1990, the pendulum swung 'nervously' from left to right, featuring an amalgamation of ideas rooted in cognitive and moral psychology (Jean Piaget and Lawrence Kohlberg), on the one hand, and, on the other hand, rational ideas rooted in behaviourist psychology (Ralph Tyler, Benjamin Bloom and Hilda Taba). For a whole decade, there lasted a sharp debate about public education. Finally, a new national curriculum was issued in 1989, featuring an intense learner-centred ideology and familiar pedagogical ideas from the progressive era. Thus, the 1989 curriculum featured what was then labelled as 'the new progressivism' (cf. Ravitch 1983). It was open-ended, advocating that boundaries between traditional subjects should be 'blotted out' (MEC 1989, p. 32) and that teaching, learning and assessment should reflect the idea of a 'whole child development'.

The old criticism against centralised examinations thus continued in the 1980s and 1890s, not least because they had been conducted as norm-referenced from 1977 to 1983. These centrally governed examinations received the term 'Samræmd próf' and later on 'Samræmd könnunarpróf', where 'samræmd' means 'coordinated' or 'centralised'. A system of relative grading was developed where the top 7% received A, 24% B, 38% C, 24% D and 7% received E. Because of entry requirements for secondary school, almost one-third of the student population received the message that they were not qualified for secondary education. The norm-referenced testing system was widely rejected by educators and was abolished, but it has in part

prevailed, although its interpretation and application have changed and the purpose is increasingly formative.

Despite a short back-to-basics period at the beginning of the new century focusing on detailed learning objectives and more centralised tests (MESC 1999), there have been no entry requirements for secondary schooling since 2002. Formative examinations (i. könnunarpróf) are first and foremost meant as supporting tools for teachers and their students providing information about strengths and weaknesses. In 2007, such formative examinations were finally presented by educational authorities as the only official assessment instruments to be used and since then no centralised achievement examinations have been used as high-stakes summative judgement about learning outcomes in Icelandic compulsory schools.

There was in increasing demand that enacted curricula in schools should receive increased attention with respect to assessment practices. Furthermore, it was suggested that teachers and schools should be responsible for both summative and formative assessment. Therefore, teachers should be provided with professional support to develop their assessment practices. Consequently, situated classroom assessment received increased attention and new conceptions began to emerge, such as authentic assessment, performance-based assessment, self-assessment, intrinsic motivation, metacognition, and last but not least, an 'old wine in new bottle' 'feedback'. Additionally, new assessment tools were introduced, such as rubrics, rating scales and porfolios.

But surprisingly, a quite different perspective caught the attention of education authorities at the turn of the century. As the emphasis on classroom assessment was gaining momentum, Icelandic authorities decided to take part in large-scale international studies of achievement such as IEA's first TIMSS study in 1995 and later OECD's PISA programme. Generally, the results of these studies of achievement have indicated a declining trend regarding achievement of Icelandic students in literacy, mathematics and science. Furthermore, reports imply that there has been a fall in the number of Icelandic students at higher proficiency levels of PISA and a rise in the number of students at lower proficiency levels.

Since the current national curriculum came into force in 2011 (MESC 2014), teachers have become increasingly responsible for assessment:

> Emphasis should be on formative assessment where pupils regularly consider their education with their teachers in order to attain their own educational goals and decide where to head. Criteria, on which the assessment is based, have to be absolutely clear to pupils. (MESC 2014, p. 26)

Furthermore, teachers have to cultivate a system of assessment criteria related to a scale (A, B+, B, C+, C, D) where A means exceptional competence, B stands for good competence, C for passable competence, and D for competence that does not reach the standard described in C. Most pupils are expected to have reached B or above by the end of compulsory education. And teachers are still reminded of their responsibility:

> In the final assessment it is of fundamental importance that teachers … make sure that the assessment is based on reliable data and that they use a variety of methods to acquire data,

> in order to give pupils, their parents and the school as clear information as possible on the pupils' status. Thus teachers can gain better insight into the studies of each pupil. For an accurate conclusion, such as from conversations or on-site inspection, it may be relevant for teachers to cooperate when they consider the data that the assessment is based on and to use precise criteria. (MESC 2014, p. 92)

The importance of teacher collaboration as maintained in the last sentence above was certainly relevant and appropriate. It entails what has been called 'moderation', that is, systematic collaboration in organising learning, and benchmarking judgements about student achievement. Research indicates that sharing common knowledge about learning outcomes and levels of achievement enhances reliability, validity and fairness regarding achievement decisions (cf. Little et al. 2003).

Relevant Research Findings

Research findings confirm that since the current national curriculum came into force, teachers and schools do need professional support when assessing student learning, both regarding theoretical issues and praxis. According to some recent findings many interrelated issues are worthy of note. Four of them are reviewed here.

First, teachers seem to face difficulties when the issue is assessing the process of learning rather than assessing what has been taught (Sigþórsson 2008; Þórólfsson et al. 2011). In other words content coverage and assessment of what has been taught seem to receive more approval than assessing learning and what has been learned. As an example a majority of participants in Sigþórsson's study (2008) admitted they were typical transmitters of knowledge relying on school books and other written resources and accordingly assessed students'knowledge and skills. Science teachers in the same study observed that proper assessment of learning was problematic; most participants were convinced that they would practice different teaching and assessments if the system allowed it, and they…

> … justified their way of teaching and how it differed from what they preferred primarily by the quantity of content that they had to cover and how it required teaching methods that enabled them to cover more content in a shorter time. (Sigþórsson 2008, p. 145)

Most intriguing was the fact that the science teachers maintained that there was not enough time and resources for hands-on learning and experiments (Sigþórsson 2008); class schedules did not allow such methods, which relates to the second issue.

The second issue concerns arranging proper conditions to assess complex and wide ranging competences such as critical thinking, problem solving, collaboration, and applying knowledge to new contexts:

> The change means that now wide ranging competences need to be assessed, and how the pupil uses knowledge and skills, not merely how good he or she is at reciting facts and remembering things by heart. A lower-secondary school principal described the changes in this way: 'It's like changing a flat tire, you need to be able to execute it, not just recite orally how to do it.' (MESC 2016)

When teachers and administrators were interviewed about assessing how pupils applied knowledge and skills, there was an agreement that informal and authentic assessment was needed, though not always easy to implement:

> We are not saying that they need to learn directly about Europe, Asia and for example rivers in Russia. Instead they need to show that they are able to read geographical maps and understand figures, graphs and tables about climate, vegetation, and such things. Thus assessment is more you know, we try to work with knowledge in class and the assessment is more about how they apply what they have hopefully learned previously. (Pétursdóttir 2018, interview with social science teacher)

The third issue of concern has to do with knowledge and skills regarding formative assessment. According to specialists such assessment is certainly not an easy job (Black and Wiliam 1998b; Leahy et al. 2005; Heritage 2010). Some teachers contend (Sigþórsson 2008) that it mainly involves regular testing during an ongoing course of instruction for the purpose of improving instruction, which is in fact a valid purpose. But formative assessment embodies a great deal of more complex teacher–student interactions and also student–student and teacher–teacher interactions. It features a process that takes place during learning and instruction where both students and teachers are active participants, 'sharing learning goals and understanding how their learning is progressing, what next steps they need to take, and how to take them' (Heritage 2010). Furthermore, it has to do with metacognition and pupils' awareness and understanding of their own thinking.

Two Icelandic studies (Pálsdóttir 2006; Þórólfsson et al. 2011) suggest that formative assessment appears as more rhetorical than real praxis. Pálsdóttir's study (2006) indicates that many schools lack clear strategies regarding assessment, especially formative assessment. Participants stated that in their schools there was a lot of discussion and work being done to develop assessment, and 'self-assessment, portfolio assessment, and peer-assessment were considered useful assessment methods' (p. 105) but they did not sense real emphasis on using them. Þórólfsson et al. (2011) found that discourse indicated focus on performance-based assessments, portfolios and authentic assessment, but 'real practice seems to endorse an academic school curriculum to a considerable extent, setting standards for students and using tests as a motivation for pupils to learn the curriculum and teachers to teach it' (p. 120).

The fourth issue concerns the transition from statistics and number grades to qualitative evaluation and letter grades. A key concept reflecting this transition is 'competence' referring to a wide range of cognitive, physical and attitudinal abilities that are supposed to be 'evaluated' by teachers not 'measured'. Consequently, in addition to knowledge and practical skills, abilities such as solving problems and organising and interpreting information are to be assessed. Studies (Pétursdóttir 2018; Þórólfsson 2017) indicate that the time lag until the new system will gain full execution may become substantial. Of those responsible for the new system in their schools (mostly administrators) in school year 2016–2017, almost two-thirds agreed or strongly agreed that their schools were insufficiently prepared for matching assessed learning outcomes with the criteria based on letter grades as stipulated in the national curriculum (Þórólfsson 2017).

Discussion

In conclusion, this historical overview demonstrates that assessment in education is an enormous issue encompassing numerous important problems and questions that educators need to consider according to context. What is the purpose of the assessment? What should be assessed? How? By whom? When? Where? How will the results (data) be interpreted? How will the results be presented and used and for what purposes?

Central professionals that these questions weigh on are teachers, who need to be well informed regarding assessment, both theoretically and empirically. Teachers need to be familiar with research and theories and be prepared to discuss with parents, students, colleagues and other professionals about the different purposes of assessment and methodology. Furthermore, they are obliged to possess knowledge of basic concepts such as validity, reliability, criteria, relative grading, and norm-referenced versus domain-referenced evaluation systems. According to law and the current national curriculum, Icelandic teachers are most responsible for reliable and valid assessment so it concerns their professional identity.

As explained above, the pendulum has swung regularly from an emphasis on measuring learning outcomes (products) to assessing the process of learning. Education and assessment have in fact reflected an amalgamation of different ideologies. Michael Schiro (2008) identified four such ideologies, scholar academic ideology, social efficiency ideology, learner-centred ideology, and social reconstruction ideology. The emphasis on measuring learning outcomes relates more to the first two and an emphasis on learning and assessment as process relates more to the last two. But as Schiro (2008) indicated, all such ideologies represent ideals abstracted from reality, not reality itself. Hence, we may experience ideas that seem real parts of the enacted curriculum, but when observed closer turn out to be more rhetorical. According to recent research in Iceland this seems to apply to formative assessment in some instances.

International comparative studies of achievement such as TIMSS and PIRLS organised by the International Association for Educational Achievement (IEA) and PISA organised by OECD have an interesting role regarding such ideologies. As stated by Schiro (2008) the social efficiency ideology aims at providing knowledge that promotes the ability to function in society, viewing learning and teaching as a process by which behaviour is shaped, and assessment as a means to confirm how well they are prepared (shaped) to function as citizens. Learners are like raw materials to be shaped according to particular objectives.

By and large, PISA embodies similar ideology, that is, social efficiency. It examines not just what students know in science, reading and mathematics, but also what they can do in real life with what they have learned. Iceland has taken part in PISA since it started in 2000. Therefore, it must be essential to observe its role and influences, because PISA is not a typical academic research enterprise: ‘It is meant to provide results to be used in the shaping of future policies … PISA concepts, ideology, values and not least the results and the rankings, shape international educational

policies and also influence national policies in most of the participating countries' (Sjøberg 2007, p. 203). Svein Sjøberg (2007, 2018) has drawn attention to some debatable features of PISA, for instance, how results are statistically reported as simple ranking in league tables, drawing attention away from more significant factors and data. Sjøberg has also identified that a written test in science can hardly measure locally situated competencies, for example, those acquired on excursions, through inquiry learning, or in experimental work. His criticism also sheds light on problems related to reliability and validity:

> … young learners in different countries and cultures may vary in the way they behave in the PISA test situation. I claim that in many modern societies, several students are unwilling to give their best performance if they find the PISA items long, unreadable, unrealistic and boring, in particular if bad test results have no negative consequence for them. (Sjøberg 2007, p. 203)

Finally, I want to re-emphasise the significance of teacher moderation. Systematic collaboration in organising learning and benchmarking judgements about student achievement is of most importance according to the current national curriculum. Networking teachers is bound to be beneficial, whether the issue is education ideologies, assessment policies, interpreting and using PISA data, or discussing assessment criteria related to wide-ranging learning outcomes and a new marking system featuring letters (A, B+, B, C+, C og D).

References

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–73.

Black, P. og Wiliam, D. (1998b). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–149.

Gíslason, M. (1977). *Kvällsvaka: en isländsk kulturtradition belyst genom studier i bondebefolkningens vardagsliv och miljö under hälften af 1800-tallet och början af 1900-tallet*. Uppsala: Studia Ethnologica Uppsaliensia.

Gunnarsson, K. J. (1963, November 13). Skólamál [School issues]. *Morgunblaðið*, pp. 15–16.

Guttormsson, L. (1981). Läskundighet och folkbildning före folkskoleväsenet. *Nordisk kulturhisoria, Nordiska hilstorikerötet i Jyväskyla i 1981*. Jyväskyla: Yliopisto.

Guttormsson, L. (2008). Fræðsluhefðin: kirkjuleg heimafræðsla. In L. Guttormsson (Ed.), *Alþýðufræðsla á Íslandi 1880–2007* (Public education in Iceland 1880–2007) (pp. 21–35). Reykjavík: Háskólaútgáfan.

Harlen, W. (2008). Teachers'summative practices and assessment for learning – Tensions and synergies. In W. Harlen (Ed.), *Student assessment and testing* (pp. 292–308). London: Sage.

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Thousand Oaks: Corwin.

Hjörvar, H. (1921). Nýjar prófaaðferðir (New examination methods). *Skólablaðið, March 1921*, pp. 30–32, and *April 1921*, pp. 45–46.

Lagemann, E. C. (1989). The plural worlds of educational research. *History of Education Quarterly, 29*(2), 183–214.

Law on the Comprensive Primary School, 1974. Lög um grunnskóla nr. 63, 1974. Reykjavík: The Althing.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment that keeps learning on track minute-by-minute, day-by-day. *Educational Leadership, 63*(3), 19–24.

Little, J. W., et al. (2003). Looking at student work for teacher learning, teacher community and school reform. *Phi Delta Kappan, 85*(3), 185–192.

ME (Ministry of Education, Menntamálaráðuneytið). (1979). *Um námsmat (Bæklingur nr. 29)* (About assessment (Pamphlet nr. 29)). Reykjavík: Author.

MEC (Ministry of Education and Culture, Menntamálaráðuneytið). (1989). *Aðalnámskrá grunnskóla* (National curriculum for compulsory schools). Reykjavík: Author.

MERC (Ministry of Education, Science and Culture, Menntamálaráðuneytið). (2016). *Nýir tímar: Aukin áhersla á hæfni og læsi í grunnskólum*. Reykjavík: Author.

MESC (Ministry of Education, Science and Culture, Menntamálaráðuneytið). (1999). *Aðalnámskrá grunnskóla* (National curriculum for compulsory schools). Reykjavík: Author.

MESC (Ministry of Education, Science and Culture, Menntamálaráðuneytið). (2014). *National curriculum for compulsory schools* in English. Reykjavík: Author.

Pálsdóttir, E. I. (2006). *Námsmat í höndum kennara* (Assessment in the hands of teachers). MEd-dissertation. Akureyri: University of Akureyri.

Pétursdóttir, D. R. (2018). *Námsmat á skilum skólastiga: Tilgangur, fyrirkomulag og nýtt matskerfi* (Assessment and the transition from compulsory education to secondary education: Purpose, nature and organisation of assessment between the two school levels). MEd-dissertation. Reykjavík: University of Iceland.

Proppé, Ó. J. (1983). *A dialectical perspective on evaluation as evolution: A critical view of assessment in Icelandic schools*. Reykjavík: Bóksala stúdenta.

Ravitch, D. (1983). *The troubled crusade*. New York: Basic Books.

Schiro, M. (2008). *Curriculum theory: Conflicting visions and enduring concerns*. Los Angeles: SAGE.

Sigþórsson, R. (2008). *Mat í þágu náms eða nám í þágu mats: Samræmd próf, kennsluhugmyndir kennara, kennsla og nám í náttúrufræði og íslensku í fjórum íslenskum grunnskólum* (Assessment for learning or learning for assessment). PhD dissertation. Reykjavík: Kennaraháskóli Íslands.

Sjøberg, S. (2007). PISA and "real life challenges": Mission impossible? In S. T. Hopmann, G. Brinek, & M. Retzl (Eds.), *PISA zufolge PISA* (PISA according to PISA) (pp. 203–224). Berlin: LIT Verlag.

Sjøberg, S. (2018). The power and paradoxes of PISA: Should inquiry-based science education be sacrificed to climb on the rankings? *NorDiNa, 14*(2), 186–202.

Thorlacius, S. (1932). *Skólamál: Einkunnagjafir og fleira* (School issues: grading etc.) (pp. 1–24). Reykjavík: Prentsmiðjan Acta.

Þórólfsson, M. Þ. (2017). *The new assessment system* (Report on changes of assessment practices in Icelandic compulsory schools (Draft)). Reykjavík: Bóksala stúdenta.

Þórólfsson, M., Sigurgeirsson, I., & Karlsdóttir, J. (2011). Námsmat í náttúrufræði: Hvað má lesa úr skólanámskrám grunnskóla? (Assessment in science education: what do school curricula emphasise?). *Uppeldi og menntun, 20*(1), 99–121.

Walker, D. F., & Soltis, J. F. (2009). *Curriculum and aims*. New York: Teachers College Press.

Chapter 15
The Long March Towards School and Student Assessment in Italy

Rosalia Castellano and Sergio Longobardi

The assessment culture has had considerable difficulty in permeating the Italian school system, and the themes of school evaluation have entered the Italian political agenda only in the last 15 years, although the Italian participation in international student assessments such as PISA, TIMSS and PIRSL has always been significant. A common denominator of the results obtained by these international surveys concerns the fairness, in territorial terms, of the Italian school system. A deeper analysis of IEA and OECD data, confirmed also by the Italian National System of Evaluation (SNV), emphasizes the presence of a skill gap between students from central-northern regions (more developed) and those from southern regions (less developed). The literacy divide is the real challenge for Italian policy makers, but its solution seems to be very far, although in the last years, an ambitious set of reforms regarding the educational system (The Good School Act, Law 107/15) and the labour market (The Industry 4.0 National Plan and Jobs Act) has been launched. These policies aimed to improve the competences of Italian students and to strengthen the linkages between the education system and the world of work, but their results will only be assessed in the long term.

The Participation of Italy in the IEA and OECD Assessments

∗∗∗In the last few years, interest in the topics of student evaluation, both from an international comparative perspective and in terms of within-country analysis, has grown significantly. Year by year, the main international student assessments, such as TIMSS (Trends in International Mathematics and Science Study), PIRLS

R. Castellano · S. Longobardi (✉)
University of Naples "Parthenope", Naples, Italy
e-mail: sergio.longobardi@uniparthenope.it

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_15

(Progress in International Reading Literacy Study), and PISA (Programme for International Student Assessment), and their country rankings have become a matter of debate not only for the stakeholders of the educational sector but more generally for the general public.

In Italy, the themes of school evaluation have entered the Italian political agenda only in the last 15 years, although the Italian participation in international student surveys has a long and solid tradition starting from 1970 (Table 15.1).

Indeed, in the 1970s, Italy, along with a few other countries, participated in the first experimental study performed by the International Association for the

Table 15.1 Italy's participation in international student assessments

Years	IEA surveys	OECD surveys
1970–1980	Six subject survey (1970–1971)	
1981–1990	SIMS (Second International Mathematics Study); SISS (Second International Science Study); Written Composition Study; PPP (Pre-Primary Project); COMPED (Computers in Education Study)	
1991–1995	PPP (Pre-Primary Project); COMPED (Computers in Education Study); TIMSS (Trends in International Mathematics and Science Study) 1995; LES 1995 (Language Education Study)	
1996–2000	PPP (Pre-Primary Project); LES (Language Education study); TIMSS-R 1999; CIVED (Civic Education Study); SITES 1998–99 (Second Information Technology in Education Study – Module 1)	SIALS 1996 (Second International Adult Literacy Survey); PISA 2000 (Programme for International Student Assessment)
2001–2005	PIRLS 2001 (Progress in international Reading Literacy study); SITES 2001 (Second Information Technology in Education Study – Module 2) TIMSS 2003	PISA 2003; ALL (Adult Literacy and Life skills)
2006–2010	PIRLS 2006; SITES 2006; TIMSS 2007; TIMSS advanced 2008; ICCS 2009 (International Civic and Citizenship Study); TEDS-M 2008 (Teacher Education and Development Study in Mathematics)	PISA 2006; PISA 2009
2011–2015	PIRLS 2011; TIMSS 2011; TIMSS 2015; TIMSS advanced 2015	PISA 2012; TALIS 2013 (Teaching and Learning International Survey); PISA 2015; PIAAC 2011/12 (Programme for the International Assessment of Adult Competencies)
2016–2020	PIRLS 2016; ECES 2016 (Early Childhood Education Study); ICCS 2016; ICILS 2018 (International Computer and Information Literacy Study)	PISA 2018; TALIS 2018

Source: Elaboration of authors on information gathered from the INVALSI and ISFOL websites (www.invalsi.it; www.isfol.it) and Damiani (2016)

Evaluation of Educational Achievement (IEA) regarding the assessment of civic and citizenship education (Six Subject Survey). Over the years, the Italian engagement and participation in IEA surveys regarding evaluation of the students in terms of both mathematics, science and reading and topics related to civic education has been ongoing.

Italy has also made a considerable effort to participate in the international survey performed by the Organisation for Economic Co-operation and Development (OECD). Italian 15-year-old students were evaluated in all editions of PISA (since 2000) both in the main domains analysed by the OECD (Reading, Mathematics and Science) and in the optional ones, such as those dedicated to financial literacy (in the 2012 and 2015 editions). In addition, Italy has actively participated in surveys concerning the teaching staff (Teaching and Learning International Survey –TALIS- 2015 and 2018) and surveys concerning the skills of the adult population, such as the ALL (Adult Literacy and Life skill) and PIAAC (Programme for the International Assessment of Adult Competencies) projects.

The Birth of the National Assessment System in Italy Between Obstacles and Delays

In Italy, there are about 8.5 millions students, attending 8600 schools, and about 872,000 people are employed as tenured teachers (year 2018/19; source: Ministry of Education, www.miur.it). The Italian school system includes: primary, lower secondary, upper secondary and higher education. The primary school starts at 6 years of age and lasts until a student is 11 years old. Students face a final exam at the end of 5-years cycle of elementary school to attend lower secondary school. At the end of lower secondary school (grade 8), students must pass another exam to gain access to upper secondary school (lasting 5 years). There are several different tracks available from grade 9 onwards and students decide (through a self-selection mechanism) which to attend choosing between four main types: Licei (schools with an academic focus that mainly cover humanities and scientific fields); art, foreign language or teacher-training schools, with an academic curriculum but a vocational orientation; Technical institutes and Professional schools. Despite the freedom of choice, in reality, family background matters a lot, "*there has always been a clear hierarchy in terms of prestige, quality of teaching and probability of enrolling at university, with the licei at the top and the Professional Institute at the bottom*" (Ballarino and Panichella 2016). At the end of upper secondary education, students must pass a formal examination (the "State Maturity exam") that allows them to access to universities or enter the world of work. In Italy, there are private schools for all levels of education; private schools account for almost 8% of the system (although they are attended by less than 5% of Italian students), and are periodically accredited by the Ministry of Education that provides state funds if they follow the same guidelines as state public schools in terms of curriculum, personnel, and management. The state has exclusive jurisdiction with regard to the general organisation

of the education system (e.g. minimum standards of education, school staff, quality assurance, public financial resources). The Ministry of Education, University and Research (MIUR) is responsible for the general administration of education at national level. At territorial level, Regional authorities have joint responsibility with the State related to some aspects of education system (definition of school calendar; distribution of schools in their territory) and have exclusive legislative competence in the organisation of the regional vocational education and training system. Schools have limited autonomy, they can define curricula, organise school time, establish learning methods, and compose the classrooms.

Italy's participation in numerous international large-scale assessments has not been accompanied by the introduction of a national evaluation system for schools and students or by the attention of policy makers to these issues. Until the end of the 1990s, the issues of student and school evaluation and the dissemination of their results have been "reserved" to a narrow circle of researchers and academicians. With the start of the new millennium, there has been a gradual inversion of course that has involved a series of initiatives on the legislative level that have begun to give a certain emphasis to the themes of school evaluation in Italy, too. The reasons for this change are to be found in various factors that, both singularly and interacting with each other, led to the affirmation of the scholastic evaluation culture also in Italy. From a within-country perspective, the greater autonomy granted to schools, with various legislative provisions in 1999 and 2000, generated a need for school accountability, i.e., the need to evaluate schools' work and make their choices and results achieved transparent. From an international point of view, the European Union increased its requests to Italy to equip itself with a system of scholastic evaluation in the same manner as the main European countries. In the early 2000s, the dissemination of PISA results not only brought media attention to the fundamental importance of students' skills for a conscious role as citizens in modern society but, at the same time, also highlighted a significant gap in the skills of Italian students compared to those of other developed countries.

The new focus on the themes of school evaluation has been accompanied by the implementation of a national system for assessing students' skills. Between 2001 and 2003, the Italian National Institute for the Evaluation of the Education and Training System (known by the Italian acronym of INVALSI)[1] performed first its experimental evaluation studies, called "Pilot Projects" ("Progetti Pilota"), which involved 2,800,000 students from the primary (grade 4) and secondary (grades 6, 9 and 11) schools -in this exploratory phase, school participation was on a voluntary basis-. Despite the low-stakes nature of these tests, several studies (Quintano et al. 2009; Bertoni et al. 2013) emphasized that teachers, especially those in the southern

[1] The INVALSI institution dates back to 1999 (Presidential Decree no.275 and Decree Law no.258). The INVALSI has scientific, organisational and financial autonomy but is subject to supervision by the Ministry of Education, University and Research (MIUR). Its task is to promote the improvement of educational attainment using national and international evaluations and to contribute to the development and growth of the Italian educational system.

regions of the country, engage in opportunistic behaviour (teacher cheating)[2] to improve their pupils' test results and consequently make the results of these experimental studies unreliable. Although the results were not encouraging, the pilot projects emphasize the difficulties associated with the creation of an evaluation system from scratch. In particular, it was clear that a statistically reliable assessment system required considerable investment in terms of human and financial resources and a need to motivate and involve the teaching staff of the schools that hindered any external evaluation process. Since 2004, numerous legislative provisions (Legislative Decree no. 286/2004, Law no. 296/2006, Decret Law no. 147/2007, Law no. 176/2007) pushed INVALSI to develop models and methodologies for students' assessment. The numerous efforts undertaken by INVALSI enabled performing during the 2008–2009 school year the first Italian national standardized tests of students' skills conducted on a national scale at several grades of the primary (grades 2 and 5) and lower secondary schools (grades 6 and 8).

The main novelties of the evaluation system were the census and obligatory nature of the tests, the incidence of the INVALSI tests on students' final marks of 8th graders students, the return of the results to the schools accompanied by regional and national benchmarks and the implementation of statistical algorithms to mitigate the impact of cheating on data quality (Quintano et al. 2009). Subsequently, several legislative provisions have made some changes to the modalities and objectives of the INVALSI test until 2013, when the Presidential Decree no. 80/2013 led to the new National Evaluation System (Sistema Nazionale di Valutazione, SNV).

The "New" Italian National Evaluation System (SNV)

The aim of the Italian SNV is to assess the efficiency and effectiveness of the national education system in order to support educational policies and to ensure the quality of education provision. The SNV is based on three pillars: external assesment of students, school self-evaluation; internal evaluation of teachers. Three main institutions are involved in the implementation of the system: (1) the INVALSI, which manages the national assessment of students outcomes and takes part in international surveys; (2) the INDIRE (National Institute of Documentation, Innovation and Research in Education), which supports the improvement and innovation pro-

[2] As highlighted by Longobardi et al. (2018), "the reasons for this behaviour are the teachers' perception that they and their school can be evaluated on the basis of INVALSI test in addition to the cultural component because teacher cheating occurs more frequently in areas that display low values of several measures of social capital (Paccagnella and Sestito 2014)". In order to minimize illegal and opportunistic behaviour, INVALSI has implemented, since the first wave of the SNV program, a series of measures both to prevent the phenomenon of cheating during the administration of the test (ex-ante) and to mitigate its impact on data quality after their collection (ex-post). With regard to ex-post measures, INVALSI has adopted the statistical procedures, developed by Quintano et al. (2009) and Longobardi et al. (2018), that allows, through a fuzzy clustering method, to detect student classes with high cheating probability and to correct their average score.

cess, the continuing professional development of school staff and documentation and research in education; and (3) a body of autonomous and independent inspectors designed by Ministry of Education (MIUR) and INVALSI.

The external evaluation of students is performed by INVALSI, which administers and proctors student tests to evaluate their knowledge and competences. Actually, in the 2017–2018 school year, all students enrolled in the second and fifth grades of primary school, in the 8 grade (lower secondary school) and in the grades 10 and 13 (upper secondary school) were required to take three tests: one in reading, one in mathematics and one in English (only for the students in grades 5 – the end of primary school – and 13 – the end of upper secondary school). The INVALSI tests in 2018 involved more than 1,100,000 students in primary schools (grades 2 and 5), approximately 570,000 students in lower secondary schools (grade 8), and approximately 550,000 students of upper secondary schools (grades 10 and 13). In addition, INVALSI administered student, teacher and family questionnaires to investigate students' socio-economic characteristics (parental education, parental wealth, home educational resources, and school mid-year marks), students' feelings and motivation and other elements useful for the evaluation of the system.

Regarding evaluation of schools, the SNV establishes that schools have to draw up a self-assessment report (Rapporto di Auto Valutazione, RAV) focused on three relevant school dimensions (context and resources, outcomes, and processes) by using and analysing the data collected by the Ministry of Education, INVALSI and the schools themselves. In addition, a random sample of schools are evaluated by an external team of inspectors (Nucleo di Valutazione Esterna, NEV) on the basis of programmes, benchmarks and protocols developed by INVALSI. At the end of evaluation, the inspectors write a final evaluation report (Rapporto di Valutazione Esterna, RVE), which is delivered to the schools.

Finally, the INDIRE takes care of supporting schools in drafting a plan of improvement actions (Piano of action and improvement). The plan highlights the main problems of the school and suggests the actions to be taken to overcome them. It focuses on four main sections: (1) choice of objectives based on the self-assessment report (RAV); (2) design of the most appropriate actions to achieve the chosen objectives; (3) planning of the process objectives; and (4) evaluation, sharing and dissemination of the results.

The evaluations of students and schools were complemented by the teacher evaluation introduced by the Law 107/2015, known as 'The good school'. The teacher evaluation is an internal process managed by school principals, who evaluate teachers according to criteria established by a Committee for the Evaluation of Teachers. The members of the Committee (chaired by the school manager) are three teachers of the school, two representatives of parents (one representative of students and one of parents for upper secondary schools) and one external member chosen by the regional school office.

The teacher appraisal evaluates the quality of teaching and the teacher's contribution to the improvement of the school in terms of student's competences, didactic and methodological innovation, research and dissemination of good practices. At the same time, teachers' assignments for organizational and management tasks are

considered. Teachers with positive evaluations receive financial bonuses at the end of school year.

The "Literacy Divide": Performance Gaps Among Italian Regions

Italy's constant participation in various international surveys has highlighted some critical aspects of the Italian educational system. Already the first edition of PISA (2000) underlined a delay, in terms of skills, of Italian students compared to those of other OECD countries. The subsequent PISA rounds have confirmed the low position of Italy in international PISA rankings and, although some improvements have been observed in the last edition (2015), Italian students still lag behind peers in reading and science, especially compared to students from countries with similar levels of socio-economic development.

The Italian results of PISA assume an additional value if they are read together with those from the IEA surveys. From an international comparative point of view, the results of TIMSS and PIRLS show that the Italian system is able to provide a considerable amount of skills to students in the first years of schooling, as indicated by the high standing of Italy in the IEA international rankings. The recent TIMSS 2015 data show that Italian fourth grade students achieve an average score of 507 points in mathematics and 516 points in science (with an international average of 500), whereas PIRLS 2016 shows that the average performance of the students themselves (quarter year) in reading is equal to 548 points, which is statistically significantly higher than both the European and international averages.

Focusing the analysis on the eighth grade of schooling, TIMSS data draws a different picture: it seems that the initial advantage of Italian students is lost during the following school years, in particular in the transition from primary to lower secondary school. Indeed, the results of the Italian students in the eighth year are significantly lower than the international averages in both mathematics and science. The poor performance of Italian students in the international rankings of PISA and TIMSS (grade 8) is a sort of "tip of the iceberg" and is a "symptom" of a more complex problem concerning the fairness, in territorial terms, of the Italian school system.

A deeper analysis of the IEA and OECD data emphasizes the presence of a skill gap between students from central-northern regions (more developed) and those from southern regions (less developed). This difference is remarkable, especially given the centralized nature of the Italian educational system and the low autonomy of Italian schools, which are constrained by their lack of power in choosing their teachers or managing the budget for their tenured staff. In this perspective, PISA data show that the very poor performance of Italian students is due to significant territorial differences within the country. Indeed, 15-year-old students in the Italian southern regions performed very low in each assessment area, which contributed to

Italy's (poor) standing in international comparisons. For each PISA cycle, the average score differs strongly among the northern and southern regions, and these marked differences cause a wide north-south divide, which is called the "literacy divide" (Quintano et al. 2012).

Panel A of Fig. 15.1 reports the differences between Italian macro areas, expressed in percentage terms compared to the national average, according to the results in mathematics of international assessments performed in 2015 (TIMSS and PISA). Focusing on the PISA data, students[3] from the northeast macro area achieve an average score of 536 points, and they do as well as top-performing students such as Koreans and Canadians. In contrast, students from southern regions consistently lag behind the others, since they exhibit an academic performance similar to the students from developing countries such as Kazakhstan and Romania. The literacy divide between northern and southern students is very large[4]; in percentage terms, the average score of students in the "South and Islands" area is 7% below the national average, on the contrary, the score of the students from the northeast area is 9% above the Italian average.

The TIMSS results shed further light on the performance gap between the Italian territorial areas. Fig. 15.1 (panel A) shows not only that the difference in mathematics is already present in the first levels of schooling (the southern fourth grade students have a 6% difference compared to the national average) but also that the gap increases significantly in the following years, when the difference between the students in the eighth year of the most- and least-developed Italian macro areas reaches 68 points (students from the northeast macro area achieve an average score of 5% above the Italian average while the southern students are 8.5% below).

Although there are many differences between the TIMSS and PISA surveys in terms of the number and composition of participating countries and the framework and purpose of the assessments (Montanaro 2008), their results are able to paint a reliable picture of the Italian school system. Indeed, this picture is fully confirmed by the national data obtained by INVALSI, which not only have a census nature but are also characterized by a high level of homogeneity in terms of both the methods of detection and the evaluation framework.

Panel B of Fig. 15.1 shows the differences of each Italian macro area from the national average (in percentage terms) computed on the basis of the data obtained by the INVALSI National Evaluation System in the 2015. Focusing on the results for second-grade students, the differences between the Italian macro areas are very low and statistically not significant, whereas for students in the fifth grade, the differences in the average results among regions increase; the southern regions achieve significantly lower results than the national average. These differences grow considerably in the transition from primary to lower secondary school and increase further

[3] The PISA results reported in fig. 1 are computed only on the sub-sample of 15-year-old Italian students who attend the 10th school grade (78.7% of the overall Italian sample).

[4] The north-south difference is equal to 81 points, it is almost equivalent to one standard deviation on the PISA international scale, which is more than 1 year of schooling.

Fig. 15.1 Differences between Italian macro areas, expressed in percentage terms compared to the national average, according to the results in mathematics of national and international assessments performed in 2015 for different grade levels. Note: The PISA results reported in fig. 1 are computed only on the sub-sample of 15-year-old Italian students who attend the 10th school grade (78.7% of the overall Italian sample). (Source: authors' elaboration on PISA 2015, TIMSS 2015 and INVALSI data)

Table 15.2 Results from regression models, according to macro-category of explanatory variables, to analyse territorial differences between Italian macro areas (dependent variable: PISA 2015 scores in science)

Covariates	Model 1	Model 2	Model 3	Model 4	Model 5
Student controls	Yes	No	No	No	Yes
School controls	No	Yes	No	No	Yes
School resources	No	No	Yes	No	Yes
Classroom and school discipline	No	No	No	Yes	Yes
Macro areas dummies reference = North East (most developed regions)					
North West	−15.37**	−14.32**	−8.30	−7.74	−7.97
Centre	−31.96***	−29.34***	−29.96***	−12.08	−18.23***
South	−55.41***	−49.67***	−49.45***	−35.90***	−49.10***
South and Islands	−78.34***	−71.42***	−72.48***	−67.75***	−65.74***
R-square	0.25	0.33	0.15	0.23	0.40

Significance level: **p<0.05, ***p<0.01. "Student control variables" include: Gender (0 = male; 1 = female); Index of economic, social and cultural status (ESCS); Immigrant status (0 = native;1 = immigrate); Grade repeat (0 = no; 1 = yes). "School control variables" include: School average of ESCS index; Location of school (village, small town, town, city, large city); Typology of school (Licei, Technical, Vocational school); Private school (0 = public, 1 = private). "School resources" include: Ratio of computers available to students (RATCMP1); Average class size (CLSIZE); Number of extracurricular activities at school; Total learning time (minutes per week), Index of staff shortage in teachers' view. "Classroom and school discipline" encompass: School average of the indices of disciplinary climate science classes (DISCLIMA); School percentage of students who had not skipped a whole school day in the 2 weeks prior to the PISA test.

between lower and upper secondary, giving rise to the literacy divide, as also highlighted by PISA.

The results of national and international assessments underline the difficulties of the Italian system in guaranteeing equal opportunity to all students regardless of their geographical area of origin.

Although these surveys complements information from the assessment of student's competences with information gathered through questionnaires administered to students, their schools and education systems, they are not able to fully explain the reasons for this performance gap.

For example, Table 15.2 shows the results of a stepwise regression analysis[5] of the Italian data from PISA 2015. These results show that the territorial gap remains high and statistically significant, even if controlling for a large set of covariates

[5] The dependent variable is student performance in science (main PISA domain in 2015) expressed by the ten PISA plausible values. Plausible values are multiple random draws from the latent student achievement. The Fay's variant of the Balanced Repeated Replication (BRR) method was used, thus accounting for PISA's complex survey design (where schools are the primary sampling units) in the estimation of sampling variances. For methodological details about the plausible values analysis and the use of replicate weights, see OECD (2009).

commonly used in the economic literature for the estimation of the Educational Production Function (EPF).

Despite the purely descriptive nature of the analysis reported in Table 15.2, it is evident that the literacy divide between northern and southern students is a complex problem that arises from many factors acting at different levels (individual, familiar, scholastic and territorial), and it is the real challenge for Italian policy makers that should be addressed at the national level.

Over the last 10 years, several studies have tried to give valid answers to explain these differences. As highlighted by Bratti et al. (2007), "the acquisition of literacy is a lifelong process taking place not just at school or through formal learning, but also through interactions with peers, colleagues and wider communities". Consequently, many factors can contribute to explain the literacy divide, including the different degree of socio-economic development between macro areas of the country both in strictly economic terms (e.g., GDP and employment rate) and in terms of social capital, quality of life, infrastructure and resources (Agasisti and Vittadini 2012; OECD 2017; Bratti et al. 2007). From this point of view, the literacy gap is likely to be lower in the first years of school, as reported by TIMSS, PIRLS and INVALSI, since families and teachers are able to protect students from the influences of the external context. Subsequently, during the adolescent years, the contribution of the family becomes less incisive, and the external environment is more able to influence the learning processes, causing worse performance in the socio-economically disadvantaged contexts that characterize the southern regions of Italy.

The Influence of International and National Assessments on Italian School Policies

In Italy, the assessment culture has had difficulty in permeating Italian schools, and consequently the birth of a national evaluation system has been troubled and delayed. Whereas some countries, such as Germany, Denmark, and Japan, experienced a PISA shock after the publication of OECD data and have implemented some political measures in response to the results of first PISA rounds, in Italy, little media attention was devoted to the PISA 2000 and 2003 results, and scarce attention was paid to the low performance of Italian students. An interesting study conducted by Breakspear (2012) analysed the policy responses to PISA 2009 of 37 participating countries and estimated an index of the potential policy impact of the PISA results. Italy is nearly at the bottom of the international ranking based on this index, and it is one of the countries where the breadth of the PISA policy impact is very low. Over the years, the attention to the evidence from international and national student assessment data has grown. Several initiatives directly related to OECD, IEA and INVALSI surveys were conducted in the period from 2007 to 2013 and

aimed to improve the poor performance of students from southern regions. For instance, the Italian Ministry of Education supported the project M@t.abel (Matematica, apprendimenti di base con e-learning) performed with EU funding (European Social Fund, or ESF, and the European Fund for Regional Development, EFRD). M@t.abel was an extensive large-scale project aimed to improve the professional development of teachers in the lower secondary and the first 2 years of upper secondary schools in four southern regions. The underlying logic of the project is inspired by the OECD perspective of learning for life. Indeed, the main idea is that students should be engaged in solving real-life problems through mathematical tools and concepts rather than learning abstract formulas and ideas. The project included teacher training, classroom experimentation through the support of tutors and experts in various disciplines and sharing of educational material and good practices. Similar initiatives were also conducted to increase Italian performance in reading and science with the projects Poseidon (materials for a linguistic education) and ISS (Teaching Experimental Science, or Insegnare Scienze Sperimentali in Italian).

In the years of 2010–2013, INVALSI implemented the 'Plan for training evaluation teams on national and international surveys' (Piano di Formazione e Informazione dei Team di Valutazione alle indagini nazionali e internazionali). This project was designed to raise awareness among teachers at southern schools regarding the topic of school evaluation and improve their knowledge about the use of national and international student assessment data to improve the educational process. In general, during the first decade of the 2000s, the attention of policy makers to the topic of student assessment and the results of international surveys have always been "indirect", as underlined by Damiani (2016): "*…references to international and national large-scale assessments refer generically to their frameworks and results, with no specifics about which cycle or what kinds of results should be addressed in order to close the student attainment gaps highlighted by the surveys*." For instance, Legislative Decree 226/2005 emphasizes the importance of an accountability process for the Italian school system and promotes the implementation of a national system of evaluation. The national guidelines for the reform of upper secondary education (regulated by DPR 87/2010, DPR 88/2010, and DPR 89/2010) reaffirm the European recommendations to create an integrated European knowledge society, and for the first time, these guidelines clearly refer to the importance of both the theoretical framework and the results of international (PISA, and IEA TIMSS) and national (INVALSI) surveys.

In recent years, there has been a new focus on these issues. Moreover, the national evaluation system (SNV) of INVALSI has achieved remarkable reliability and quality in statistical terms.

From this perspective, the literacy scores of Italian students obtained by the INVALSI SNV were inserted, by the Italian National Statistical Office (ISTAT), into the set of national indicators for the achievement of the United Nations Sustainable Development Goals (SDG) of the 2030 agenda (Goal 4, quality of

education[6]). In the period of 2015–2017, the topic of the school and students' assessment fully entered the Italian political agenda, although in a broader sense and in relation to the needs of the labour market. In these years, the Italian government launched an ambitious set of reforms that focussed on the central role of skills for economic growth, wellbeing and socio-economic inclusiveness.

The Good School reform (La Buona Scuola, 2015, Law 107/15) was introduced in 2015 to transform the Italian education system. The Good School act includes several measures related to the school autonomy, the recruitment of a significant number of new teachers, the introduction of a merit-based component of teachers' salaries and the enhancement of digital innovation and skills in schools to improve educational outcomes of Italian students. The reform establishes the centrality of students' skills as the primary result of the educational process, as theorized in the theoretical framework of OECD PISA. It is interconnected with some recent labour market reforms (the Industry 4.0 National Plan and Jobs Act) with the aim of improving student skills and strengthening the linkages between education system and the world of work.

Undoubtedly, the implementation of these laws requires sustained and lasting efforts both from school principals and teachers and from families, employers and other stakeholders (OECD 2018), and their results will only be assessed in the long term. In conclusion, the path of Italy's march towards a school and student evaluation system has been long and not without obstacles, but the road is still long, and many challenges, such as the literacy divide, remain open.

References

Agasisti, T., & Vittadini, G. (2012). Regional economic disparities as determinants of students' achievement in Italy. *Research in Applied Economics, 4*(1), 33–54.

Ballarino, G., & Panichella, N. (2016). Social stratification, secondary school tracking and university enrollment in Italy. *Contemporary Social Science, 11*(2–3), 169–182.

Bertoni, M., Brunello, G., & Rocco, L. (2013). When the cat is near the mice won't play: The effect of external examiners in Italian schools. *Journal of Public Economics, 104*, 65–77.

Bratti, M., Checchi, D., & Filippin, A. (2007). Geographical differences in Italian students' mathematical competencies: Evidence from PISA 2003. *Giornale degli Economisti e Annali di Economia, 66*(3), 299–333.

Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance* (OECD Education Working Papers, No. 71). Paris: OECD Publishing. https://doi.org/10.1787/5k9fdfqffr28-en.

Damiani, V. (2016). Large-scale assessments and educational policies in Italy. *Research Papers in Education, 31*(5), 529–541. https://doi.org/10.1080/02671522.2016.1225354.

[6]The Italian Sustainable Development Goals related to INVALSI data are Global indicator 4.1.1. (Proportion of children and young people in grades 2/3; at the end of primary; and at the end of lower secondary achieving at least a minimum proficiency level in reading and mathematics, by gender) and Global indicator 4.5.1. (Parity indices for literacy and numeracy).

Longobardi, S., Falzetti, P., & Pagliuca, M. M. (2018). Quis custiodet ipsos custodes? How to detect and correct teacher cheating in Italian student data. *Statistical Methods and Applications, 27*, 515–543. https://doi.org/10.1007/s10260-018-0426-2.

Montanaro, P. (2008). *Territorial differences of Italian students' achievement: Evidence from national and international assessments* (Bank of Italy Occasional Paper No. 14/08). Rome.

Organisation for Economic Cooperation and Development. (2009). *PISA data analysis manual: SPSS and SAS* (2nd ed.). Paris: OECD.

Organisation for Economic Cooperation and Development. (2017). *Getting skills right: Italy, getting skills right*. Paris: OECD Publishing. https://doi.org/10.1787/9789264278639-en.

Organisation for Economic Cooperation and Development. (2018). *OECD skills strategy diagnostic report: Italy 2017, OECD skills studies*. Paris: OECD Publishing. https://doi.org/10.1787/9789264298644-en.

Paccagnella, M., & Sestito, P. (2014). School cheating and social capital. *Education Economics, 22*(4), 367–388. https://doi.org/10.1080/09645292.2014.904277.

Quintano, C., Castellano, R., & Longobardi, S. (2009). A fuzzy clustering approach to improve the accuracy of Italian student data. *Statistica & Applicazioni, 7*(2), 149–171.

Quintano, C., Castellano, R., & Longobardi, S. (2012). The effects of socioeconomic background and test-taking motivation on Italian students' achievement. In A. Di Ciaccio, M. Coli, & J. M. A. Ibañez (Eds.), *Advanced statistical methods for the analysis of large data-sets* (pp. 1–484). Berlin: Springer.

Chapter 16
Large-Scale Assessments in the Norwegian Context

Henrik Galligani Ræder, Rolf Vegar Olsen, and Sigrid Blömeke

This chapter provides a brief overview of national and international large-scale assessment in Norway. Embedded in a range of assessment tools that consist of mapping tests in grades 1–4, national assessments in grades 5–9, national exams at the end of lower and upper secondary school and student surveys in grades 7–11, the international large-scale assessments (ILSAs) have a specific role. This role is described, as well as the assessment system as a whole. Norwegian results from the ILSAs are presented with a focus on long-term developments since the mid-1990s and equity as the most characteristic result regarding Norway seen from an international perspective. Finally, the benefits and limitations of the assessment system in its whole, and with its different tools, are discussed against a framework that distinguishes between educational monitoring, support for teaching and learning and certification as core functions of educational assessments. Conclusions are drawn regarding the possibilities to further develop the whole assessment system and its individual tools.

Introduction to National and International Large-Scale Assessments in Norway

This chapter discusses the large-scale assessments in place during the first 10 years of the Norwegian school system. The foundation of education in Norway is a comprehensive school system during the first 10 years, with students starting the year they turn six. More than 95% of the students enrolled in grades 1–10 attend a public school, and attending primary school (1–7) and lower secondary school (8–10) is mandatory for all children in Norway. As the placement is decided geographically

H. G. Ræder (✉) · R. V. Olsen · S. Blömeke
Centre for Educational Measurement at the University in Oslo (CEMO), Oslo, Norway
e-mail: h.g.rader@cemo.uio.no

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_16

for primary and lower secondary schools, more populated areas see stark differences between schools with regard to students' background. Furthermore, there are large differences between rural and urban areas regarding school size, as 30% of the primary schools in Norway have less than 100 students, serving about 6% of the students. The number of small schools and students attending them has however been in continual decline the last decades.

Like many other Northern European countries, the Norwegian assessment system is based on certification through a combination of teacher marks and exit exams. But it differs with formal marking first being introduced in grade 8 (i.e. at the lower secondary level). Norway combines a centralised curriculum with decentralised responsibilities. Traditionally, the Norwegian school system is rooted in a national curriculum combined with an Education Act specifying relatively detailed requirements for schools as a public service. Simultaneously, Norwegian schools are decentralised in the sense that the schools are under local municipal ownership and control. Furthermore, the fact that about 80% of students' final scorecards consists of marks set by the teachers themselves implies that assessment practices and marking, to a large degree, are decentralised to the local level.

This underlying structure has been constant for a long period of time (for a historical perspective, see Lysne 2006). However, over recent decades, the policies and practices of assessment have seen major changes. In the late 1980s, the Organisation for Economic Cooperation and Development (OECD) conducted a broadly scoped review of the Norwegian educational system. Although this report (OECD 1988) praised several features of the Norwegian school system, it raised concerns regarding a combination of many small municipalities/schools with strong local autonomy and virtually no central system for quality monitoring or inspections used for accountability. Furthermore, the OECD noted that for the same reason, the central government was not in a position to support policymaking with evidence and knowledge of the status of the educational system.

Although much debated and discussed, the report did not have an immediate effect on policymaking. However, it became a reference point for later decisions. Following the OECD report, Norway joined the major international large-scale assessments (ILSAs) in the 1990s and early 2000s. This can be interpreted as a degree of awareness for the need to have representative data at the national level to help monitor the outcomes of schooling. Participating in precursors to both the Progress in International Reading Literacy Study (PIRLS) and the Trends in International Mathematics and Science Study (TIMSS), Norway followed up by participating in all components and populations in TIMSS and has participated in every cycle of the Programme for International Student Assessment (PISA) since the start in 2000.

According to the then minister and the state secretary to the Minister of Education, the results from the first PISA survey disseminated in 2001 were a wake-up call (frequently referred to as 'the PISA shock'). That Norway, as a country with one of the largest investments in education, should perform close to the international average was not expected or acceptable. A metaphor used to communicate this shock was the comparison with Norway coming home from a winter Olympics without

any medals. Following this report, a series of initiatives to reform the Norwegian educational system were taken (Bergesen 2006). As an almost immediate response, an expert committee was put to work by the government with a mandate to suggest how the system should improve. The reports from this group were exceptional, in the sense that they formulated several specific recommendations (NOU 2002:10, 2003:16). These Green Papers had a decisive impact on educational policymaking over the next years to come. One of the central recommendations was to build a national system for monitoring the qualities of outcomes, processes and structures.

This recommendation was immediately followed up by parliamentary resolutions establishing the National Quality Assessment System (NQAS). With the aims of assessing 'the overall learning outcome with emphasis on knowledge, skills and attitudes' and 'the process quality in order to create as good learning environments as possible' (Proposition to the Parliament 2002–2003:1, Ch. 2), multiple assessment components have been developed and incorporated. In the next section, the components of the NQAS regarding student assessment are described in more detail before we present key results from the ILSAs.

Large-Scale Assessments in Norway Today

The core feature of the NQAS is an interactive database, the school portal (skoleporten.udir.no). This interactive portal allows everyone to produce reports for selected units, from the national level to the level of a specific school, containing average results for indicators of learning outcomes, processes and resources. In this chapter, we present and discuss the five most central components of large-scale students' assessment in the NQAS: the exit exams, national assessments, mapping tests, the student survey and ILSAs.

The National Exit Exams

The traditional form of assessment in Norway is exit exams at the end of each of the two secondary levels of schooling. The exams come in three different formats: written, oral and practical. The written exams are developed and marked centrally, while the oral and practical exams are developed and marked locally according to national regulations. Although there are a large number of exams, each student is selected to sit only a limited number. Their primary purpose is to assess students' mastery of school subject-specific learning outcomes, and the results are reported on the students' scorecard together with the teacher grades. Student scorecards are the most commonly used for selection criteria when students move into subsequent higher educational levels. As such, exams are high-stakes tests, but given that teacher grades dominate on the scorecard, the impact of the (relatively few) exams is less than in many other countries.

National Assessments

In 2004, a set of national large-scale assessments were introduced at several levels of primary and lower secondary education mainly with the purpose of educational monitoring at the system level, defining 'system' as all levels from the municipalities to the national. The tests were immediately met with criticism from several stakeholders. An evaluation also identified severe deficiencies with the tests (Lie et al. 2005). After a few years they were reintroduced in improved versions, and, with some minor changes since, the assessments now consist of two sets of tests administered at the beginning of 5th and 8th/9th grades (Directorate for Education and Training 2017).

The tests cover reading comprehension, English reading skills and mathematical literacy. They are low stakes for students, but the schools are held accountable through so-called result dialogues with the municipality administration (Mausethagen et al. 2018). Since 2014, the tests have had an anchor design allowing for horizontal equating and thus comparisons of trends in outcomes over years (Björnsson 2018). However, the 5th and 8th/9th grade tests are not linked, making them unsuitable for tracking the progress of students.

The Mapping Tests

In addition to these national assessments, another set of tests was introduced at around the same time. These tests were specifically designed to identify students at risk of falling behind in the first school years (Directorate for Education and Training 2018b). Tests of reading for grades 1–3 and of numeracy in grade 2 are mandatory for all schools and students. In addition, schools can voluntarily administer centrally developed mapping tests in numeracy (grades 1 and 3), English (grade 3) and ICT literacy (grade 4). From the fall of 2020, the reading test for grade 1 will no longer be mandatory (however, the practical effect of it being voluntary may be limited due to the high number of schools choosing to use the other voluntary mapping tests). The students at risk are identified by a cut-score set, approximately at the 20th percentile, based on a representative sample from the first administered test form, with the lifespan of each form approximately 5 years. The data from the mapping tests are handled and stored locally at the school, reinforcing that the tests are intended as diagnostic tools and not monitoring devices.

The Student Survey

The student survey measures dimensions of students' psychosocial learning climate (e.g. well-being, motivation, teacher support, safety, home–school cooperation) (Directorate for Education and Training 2018a). Originally introduced in 2001, this

survey was incorporated into the NQAS in 2004. The student survey is compulsory for schools to administer in grades 7, 10 and 11. Students' responses are anonymous, and they can opt out if they do not want to take part. Furthermore, schools can voluntarily administer the survey for all other grades from grades 5 to 13. Approximately 75% of students in grades 5–13 participated in the most recent surveys (Wendelborg et al. 2017).

ILSAs in Norway

Since 1995, Norway has participated in almost all cycles of the major ILSAs organised by the International Association for the Evaluation of Educational Achievement (IEA) and the OECD. This implies that samples of students, teachers and principals regularly participate in PISA, TIMSS, TIMSS Advanced, PIRLS, the International Civic and Citizenship Education Study (ICCS), the International Computer and Information Literacy Study (ICILS), the Teaching and Learning International Survey (TALIS) and the Starting Strong Survey. In the NQAS, results from these assessments are used for monitoring at the national level.

Some Key Results from ILSAs

In this section, we discuss some of the major findings from ILSAs, focusing on PISA (implemented in grade 10 in Norway), TIMSS (grades 4 and 8) and PIRLS (grade 4). These three studies give us the opportunity to highlight how the Norwegian system has changed (or not) over the two last decades – which also coincides with the period described above, in which the assessment system saw a change towards a more systematic approach to assessment as a tool for quality monitoring.

Long-Term Developments in Norwegian ILSA Results

Figure 16.1 shows an overview of the development of Norway's scores in the ILSAs mentioned. The figure should be read with a caveat: the various international studies assess different constructs, and no direct comparisons should be made between the studies. The figure does, however, illustrate how all these assessments present a reasonably coherent picture of the trend over time.

In short, the figure tells a story of decline in the first period. Students starting school in the late 1990s represent the low point, which can be seen around 2003–2006. Some elements of this decline are rather dramatic: from 1995 to 2003, there is a decrease of approximately 40 points for the cohorts participating in the TIMSS populations, which equalled roughly 1 year of schooling in both the fourth

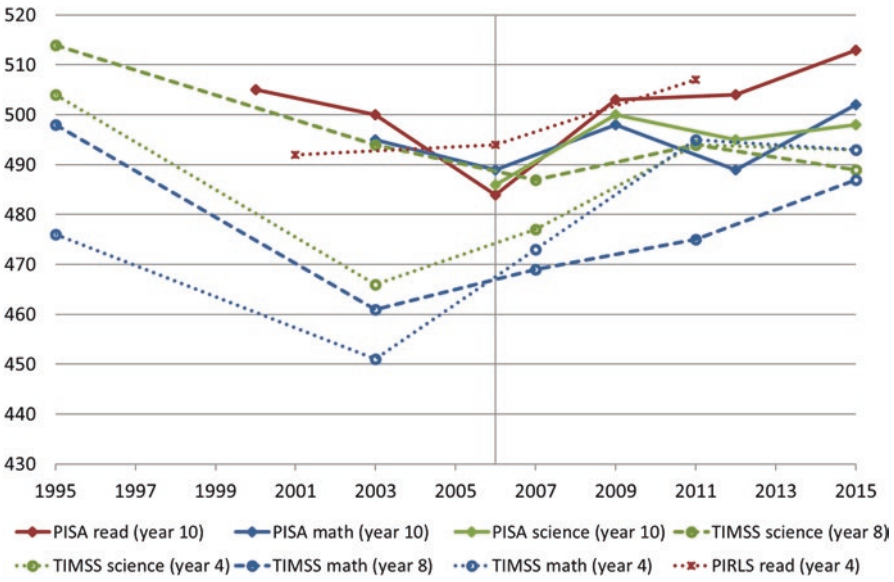

Fig. 16.1 Average scores for all populations in PISA, PIRLS and TIMSS 1995–2015. Scores for each study are represented by the originally reported scales from the studies. (Taken from Olsen and Björnsson 2018, p. 21)

and the eighth grade. In science, for example, Norway saw the second strongest decline among all participants. These results were later supported by results from the Programme for the International Assessment of Adult Competencies (PIAAC) 2012, where the age group 16–24 performed considerably worse in reading and mathematical literacy than the previous cohorts (Bjørkeng and Lagerstrøm 2014).

In the last half of the period represented in Fig. 16.1, the trend is reversed with an approximately equally large increase in scores across cohorts and domains (with eighth grade science as the most visible exception. Olsen and Blömeke (2018) analysed this increase by applying an Oaxaca-Blinder decomposition approach to the trend in mathematics in grade 8 between 2003 and 2015. The analysis established that the student composition had changed, mainly by a doubling of students with a migrant background, which should predict a decrease in the national average. This was compensated by a positive development in students' self-concept, motivation and learning environment. Nevertheless, the analysis also revealed that the increase, to a large extent, is related to factors which are not observed in the TIMSS study.

Equity in Norway

Another main feature of the Norwegian profile in the ILSAs is a high degree of equity. This can be seen from three key components of equity emphasised by Strietholt (2014): the relationship between students' socio-economic status (SES)

and performance, the distribution of performance and the proportion of students meeting minimum requirements.

The relation of SES to performance is comparably low in Norway, from an international perspective (Organisation for Economic Cooperation and Development 2016), and, in contrast to many other countries, ILSA results have not shown an increase in this relationship in recent decades (Nilsen et al. 2018). In addition, there has been a general reduction in the variance of the achievement scores in PISA, TIMSS and PIRLS. Adjusting the variance relative to the first year of both PISA and TIMSS, the proportion among schools has always been low and stable around 10%, while the overall variance, and thus the within-school variance, has decreased over the years (Nilsen et al. 2018). Finally, it should be noted that the decline in Norway's achievement from 1995 to around 2003/2006 happened along the full range of performance, whereas the following increase in scores is mostly accounted for by a shift upwards at the lower end of the distribution (Olsen and Björnsson 2018).

Gender differences in both science and mathematics have also been small or non-significant in both primary and lower secondary schools, both overall and in subdomains (e.g. Beaton 1996; Kjærnsli and Jensen 2016). However, the picture changes when reading is considered. In the PISA assessments, the gender gap has been consistently among the largest in the world, with girls outperforming boys (approximately 50 points on the PISA scale).

Public Debates Around ILSAs

Although earlier ILSAs had already revealed that Norwegian students were performing around the international average, the results from PISA 2000 caused a public 'shock' (Bergesen 2006) and put education on the agenda. The implementation of the NQAS described above can be regarded as an immediate outcome of this debate. Recent public attention on the release of ILSAs has seen positive developments: while media reporting for a long time focused on international rankings and comparisons with our neighbouring countries, media coverage has, in recent years, moved towards more nuanced considerations like equity; furthermore, at least from our perception, policymakers tend to 'cherry-pick' results less than was seemingly the case earlier (Nortvedt 2018). In addition, the critique of ILSAs, and in particular of PISA, has been present both in academic and popular media (e.g. Sjøberg 2016).

At the school level, the same tendency can be observed regarding the coverage of national LSAs. The school portal mentioned previously ensures the availability to the public of descriptive data at the school level, but the design does not initially allow for reports of ranked lists. However, with some effort, this can easily be done, and, consequently, the media has regularly published league tables, for example, as part of stereotypical stories with the narrative of 'naming, shaming and blaming' of schools (Elstad 2009). This tendency has substantially decreased in recent last years.

Discussion: Towards a More Holistic Assessment System

Norway certainly has come a long way in building a large-scale assessment system during the last 15 years, and, to a large degree, the system in its current state is able to meet the aim of assessing 'the overall learning outcome with emphasis on knowledge, skills and attitudes' (Proposition to the Parliament 2002–2003:1, Ch. 2). At the national level, this aim is achieved through the use of ILSAs and the other national assessment systems. At the local level, this aim is met in a standardised fashion, with few opportunities to adjust to specific local needs. However, the system is unable to meet the more ambitious goal of assessing 'the process quality in order to create as good learning environments as possible' (Proposition to the Parliament 2002–2003:1, Ch. 2). The current system primarily provides descriptive data from individual cross-sectional measures. To fulfil the more ambitious aim, there is a need to rethink the national assessment system, starting with a holistic framework connecting the different assessment components to each other.

A first requirement for a holistic framework would be to define the main purposes of each tool in the overall assessment system. Today, most of the assessments have multiple and simultaneous purposes (some of which are explicitly stated). It is well known that this may lead to a situation with uncontrolled 'function creeping', potentially jeopardising the validity and usefulness of the assessments (Koretz 2016). In this context, it may be helpful to distinguish between educational monitoring at the different system levels from other functions, such as support for teaching and learning or certification (Tveit and Olsen 2018).

Policymakers and stakeholders at all system levels need information about how effective the resources used in schooling are in terms of outcomes. In society, there will always be other potential allocations of these resources, and decisions regarding the level of investment in education will, therefore, constantly need to be rationalised or even defended. At the national level, sample-based studies would be sufficient for this purpose (Greaney and Kellaghan 2007). This would also lessen the burden and time used for assessment at the local level since each school would only occasionally be included in the samples. Furthermore, this approach avoids the notions of top-down control and provides data suited for research purposes.

However, the assessment system is intended to provide information on a multitude of levels. Due to the small size of many municipalities and schools in Norway, a sample-based approach would not provide actionable information to schools and school owners. In an evaluation of the NQAS principals, school owners expressed that they need data from assessments and surveys to inform local decision-making and quality development (Allerup et al. 2009), likely reflecting that the Education Act requires school owners to monitor and document the qualities of their schools.

Furthermore, feedback and support for teaching and learning are also a purpose highlighted by the NQAS for several of the assessments. This implies that data should be used to inform practices at classroom and student levels. This would require the provision of supporting material to help teachers make good use of the results, and it further points towards comprehensive assessment across the whole

cohort of students. Lastly, to support the interpretation of test results at the individual level, precision in the test scores is crucial.

Both for monitoring and for support for teaching and learning, it would be helpful to have longitudinal data making it possible to track student progression over time. A recent Green Paper in Norway highlighted the importance of the curriculum and instruction based on a clear idea of students' progress (NOU 2015:8). Since several assessments are already in place, the next logical step would be to connect these. Starting early would be a crucial aim in this context, which means that – ideally – the national assessments in grades 5 and 8/9 should be linked to the early age mapping tests. However, currently these tests are designed from very different principles and purposes, which make linking hard, if not impossible. The mapping tests are optimised to have maximum information at the cut-score (20th percentile), resulting in a highly skewed distribution and in ceiling effects. Accordingly, the scores for most students are unreliable. A possible solution for keeping the initial purpose of identifying students at risk, while at the same time providing reliable scores across the proficiency spectrum, would be to transform the mapping tests into adaptive tests – as is done, for example, in Denmark (Bundsgaard 2018).

A trial is currently being implemented by the authors of this chapter regarding linking the assessments of mathematical literacy from grades 5 to 9. These assessments are constructed from the same design principles with similar frameworks, and initial analysis looks promising regarding implementing a relatively cost-efficient design for vertically scaling the two assessments. Such a linked assessment design would also make it possible to estimate the value added of schools directly as the difference in scores between two or more time points. Furthermore, having linked national assessments would create a vital resource for studies evaluating the effects of reforms or more targeted interventions.

A broader perspective on the outcome of schooling is promoted in research and current policy documents in Norway and other countries. This include constructs such as students' motivation, perseverance and social well-being. Such measures are included in the student survey in Norway. The current system with anonymous responses is well argued for from a personal protection point of view (in compliance with, for instance, stricter regulations put into action in the European Union), and it helps ensure that students can report truthfully about their relationship with their teacher, as well as other personally sensitive issues. However, the fact that the survey is voluntary for students makes it possible that self-selection might be a source of bias. Furthermore, the implementation of the survey is not standardised, casting some doubt about the comparability of the data across schools (Wendelborg and Caspersen 2016).

Looking forward, connecting the data from the ILSAs to other data sources should be considered. Norway has an excellent base of register data tracking a broad range of variables at the individual level, such as health data, parental education, income, line of work and the housing situation of the full population. Incorporating data from the ILSAs into the national registry database would allow for anchoring the results from national assessments in an international context and would provide better measures of evaluating students' backgrounds than their self-reports in the ILSAs.

A final function of assessments is to certify a certain level of knowledge and/or skills (Tveit and Olsen 2018). The national exams serve mainly this purpose, but they are also used for local and national quality control (Mausethagen et al. 2018). Several routines are in place to ensure the quality of the exams: there is a common framework, and the tasks are developed by larger groups of expert teachers. However, little is known about the reliability or validity of the exit exams. Furthermore, the documents regulating the development and implementation of exams do not give test specifications or detailed quality criteria. This lack of knowledge about the quality of the exams as measures of students' subject-specific knowledge, skills and abilities, as well as the lack of formulations about how exams should or should not serve a range of purposes (beyond being a summative and final evaluation), may also lay open a range of unintended effects (Tveit and Olsen 2018).

As said previously, we need to note, in general, that a lack of research on national large-scale assessments exists. This lack applies to all types of traditional psychometric criteria but also to the use of outcomes by practitioners. Do they appropriately use the data for the purposes intended, as required by the current notions of validity (e.g. Kane 2013)? The status for quality assurance is nevertheless very different in Norway today compared to 30 years ago.

References

Allerup, P., Kovac, V., Kvåle, G., Langfeldt, G., & Skov, P. (2009). *Evaluering av det nasjonale kvalitetsvurderingssystemet for grunnopplæringen* [Evaluation of the national quality assessment system for primary and secondary education] (FoU Rapport, 8). Kristiansand: Agderforskning.

Beaton, A. E. (1996). *Mathematics achievement in the middle school years. IEA's third international mathematics and science study (TIMSS)*. Boston: Center for the Study of Testing, Evaluation, and Educational Policy.

Bergesen, H. O. (2006). *Kampen om kunnskapsskolen* [The struggle for a knowledge-based school]. Oslo: Universitetsforlaget.

Bjørkeng, B., & Lagerstrøm, O. (2014). *Voksnes basisferdigheter–resultater fra PIAAC* [Adult basic skills—Results from PIAAC]. *SSB Reports 2014/19*. Oslo/Kongsvinger: Statistics Norway.

Björnsson, J. K. (2018). Om lenkefeil og ekvivaleringsmetoder på nasjonale prøver: Evaluering av endring over tid [Linking-error and equating methods for the national assessments: Evaluation of changes over time]. *Acta Didactica Norge, 12*(4), 1–24.

Bundsgaard, J. (2018). Pædagogisk brug af test [Pedagogical usage of tests]. *Sakprosa, 10*(2).

Directorate for Education and Training. (2017). *Rammeverk for nasjonale prøver* [The national testing framework]. Oslo: Directorate for Education and Training.

Directorate for Education and Training. (2018a). *Elevundersøkelsen* [The pupil survey]. Oslo: Directorate for Education and Training.

Directorate for Education and Training. (2018b). *Kva er kartleggingsprøver?* [What are the mapping tests?]. Oslo: Directorate for Education and Training.

Elstad, E. (2009). Schools which are named, shamed and blamed by the media: School accountability in Norway. *Educational Assessment, Evaluation and Accountability, 21*(2), 173–189. https://doi.org/10.1007/s11092-009-9076-0.

Greaney, V., & Kellaghan, T. (2007). *Assessing national achievement levels in education*. Washington, DC: The World Bank.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kjærnsli, M., & Jensen, F. (2016). *Stø kurs. Norske elevers kompetanse i naturfag, matematikk og lesing i PISA 2015* [On track. Norwegian students' competency in science, numeracy and reading in PISA 2015]. Oslo: Universitetsforlaget.

Koretz, D. (2016, April). *Measuring postsecondary competencies: Lessons from large-scale K-12 assessments*. Paper presented at the Invited keynote address, KoKoHs (Modeling and Measuring Competencies in Higher Education) International Conference, Berlin, Germany.

Lie, S., Hopfenbeck, T., Ibsen, E., & Turmo, A. (2005). *Nasjonale prøver på ny prøve: Rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2005* [National tests being retested: Report from a committee study analysing and assessing the quality of tasks and results of national tests, spring 2005]. Oslo: Department of Teacher Education and School Research.

Lysne, A. (2006). Assessment theory and practice of students' outcomes in the Nordic countries. *Scandinavian Journal of Educational Research, 50*(3), 327–359.

Mausethagen, S., Prøitz, T. S., & Skedsmo, G. (2018). *Elevresultater. Mellom kontroll og utvikling* [Student results. Between control and development]. Bergen: Fagbokforlaget.

Nilsen, T., Björnsson, J. K., & Olsen, R. V. (2018). *Hvordan har likeverd i norsk skole endret seg de siste 20 årene?* [How has equity in Norwegian schools changed the last 20 years?]. In J. K. Björnsson & R. V. Olsen (Eds.), *Tjue år med TIMSS og PISA i Norge: Trender og nye analyser* [Twenty years with TIMSS And PISA in Norway: Trends and new analysis] (pp. 150–172). Oslo: Universitetsforlaget.

Nortvedt, G. A. (2018). Policy impact of PISA on mathematics education: The case of Norway. *European Journal of Psychology of Education, 33*(3), 427–444.

NOU (Government Official Report). (2002:10). *Første klasses fra første klasse. Forslag til rammeverk for et nasjonalt kvalitetsvurderingssystem av norsk grunnopplæring* [Proposed national quality assessment framework for primary and secondary education]. Oslo: Ministry of Education and Research.

NOU (Government Official Report). (2003:16). *I første rekke. Forsterket kvalitet i grunnopplæringen for alle* [A better education for all]. Oslo: Ministry of Education and Research.

NOU (Government Official Report). (2015:8). *Fremtidens skole. Fornyelse av fag og kompetanser* [The school of the future. Renewal of subjects and competences]. Oslo: Ministry of Education and Research.

Olsen, R. V., & Björnsson, J. K. (2018). *20 år med internasjonale skoleundersøkelser i Norge: Bakgrunn, læringspunkter og veien videre* [20 years with international large scale assessments in Norway. Background, lessons and the road ahead]. In J. K. Björnsson & R. V. Olsen (Eds.), *Tjue år med TIMSS og PISA i Norge: Trender og nye analyser* [Twenty years with TIMSS And PISA in Norway: Trends and new analysis (pp. 12–34). Oslo: Universitetsforlaget.

Olsen, R. V., & Blömeke, S. (2018). *Hva forklarer endringer i elevenes matematikkprestasjoner over tid?* [What explains the change in students' mathematics performance over time?]. In J. K. Björnsson & R. V. Olsen (Eds.), *Tjue år med TIMSS og PISA i Norge: Trender og nye analyser* [Twenty years with TIMSS And PISA in Norway: Trends and new analysis] (pp. 128–149). Oslo: Universitetsforlaget.

Organisation for Economic Cooperation and Development. (1988). *Reviews of national policies. Norway*. Paris: OECD Publishing.

Organisation for Economic Cooperation and Development. (2016). *Results (volume I): Excellence and equity in education*. Paris: OECD Publishing.

Proposition to the Parliament. (2002–2003:1). *Tillegg nr. 3 (2002–2003) for budsjetterminen 2003: Om tilleggsforslag i statsbudsjettet for 2003 under kapitler administrert av Utdannings- og forskningsdepartementet* [Amendment no. 3 (2002–2003): On supplementary proposals in the state budget for 2003 under chapters administered by the Ministry of

Education and Research]. Retrieved from https://www.regjeringen.no/no/dokumenter/stprp-nr-1-tillegg-nr-3-2002-2003-/id435850/sec2

Sjøberg, S. (2016). OECD, PISA, and globalization: The influence of the international assessment regime. In C. H. Tienken & C. A. Mullen (Eds.), *Education policy perils – Tackling the tough issues* (pp. 102–133). New York: Routledge.

Strietholt, R. (2014). Studying educational inequality: Reintroducing normative notions. In R. Strietholt, W. Bos, J.-E. Gustafsson, & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 51–58). Waxmann: Münster.

Tveit, S., & Olsen, R. V. (2018). Hvilke formål og roller har eksamen i norsk grunnopplæring? [What are the purpose and roles of exams in Norwegian primary and secondary education?]. *Acta Didactica Norge, 12*(4).

Wendelborg, C., & Caspersen, J. (2016). *Høyt presterende elevers vurdering av læringsmiljøet: Analyser av elevundersøkelsen 2013 og 2014* [High-performing students' assessment of the learning environment: Analysis of the pupil survey 2013 and 2014]. Trondheim: NTNU Samfunnsforskning.

Wendelborg, C., Røe, M., Utvær, B. K. S., & Caspersen, J. (2017). *Elevundersøkelsen 2016: Analyse av elevundersøkelsen 2016* [The pupil survey 2016]. Trondheim: NTNU Samfunnsforskning AS.

Chapter 17
International Large-Scale Assessments: Trends and Effects on the Portuguese Public Education System

João Marôco

Portugal's first major participation on an international large-scale student assessment (ILSA) was with TIMSS 1995. Portuguese students performed very poorly, and this led to the dismissal of TIMSS as a valid assessment for the Portuguese education system and the devaluation of the results. However, participation in TIMSS 1995 set in motion a sample-based external assessment in mathematics the following year whose framework was clearly inspired by TIMSS and which set the reference for the external assessments that still prevail today. The PISA 2000 survey that followed found the Portuguese students in the lowest rankings amongst the OECD countries. The TIMSS and PISA results reinforced the perception that the Portuguese curricula, teaching and assessment practices needed much improvement. ILSA and OECD suggestions were used to support and justify education policies aimed at the curricula reformulation, teaching practices, students' support programs and schools' management. In 2015, Portugal ranked for the first time above the OECD PISA average. This chapter gives a brief overview of the Portuguese record in major ILSAs and their effects on the shaping of the Portuguese educational system.

The Portuguese Education System

Universal education in Portugal has a long tradition that goes back to the early nineteenth century with the publication of the Constitutional Bill of Civil Rights that determined free primary education – focused on reading, writing, and mathematical calculations – for all Portuguese citizens (Mendonça n.d.). Up to 1910, education in Portugal was primarily provided by religious orders, namely, the

J. Marôco (✉)
ISPA-Instituto Universitário, Lisboa, Portugal
e-mail: jpmaroco@ispa.pt

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_17

Jesuits. The implementation of the Republic in October of that year overthrew the monarchy and brought the expulsion of the religious orders from Portugal and the first reform of the Portuguese education system (MEC-OEI 2003; Mendonça n.d.). Concerned with the decline of the Portuguese school system, the high illiteracy rates – about 70% of the population – and its lag in relation to most European countries, the first Republic's government set a phased large reform programme for the various levels of education (MEC-OEI 2003). Key educators of that time, like João de Deus, led the reform of the primary education system, emphasizing the importance of reading literacy at early ages (Candeias et al. 2007; MEC-OEI 2003; Mendonça n.d.)

With the military revolution of 1926 that opened the door to the fascist regime that followed, education reforms succeeded mainly with an emphasis on ideological nationalism. During the 1950s, illiteracy rates were still high – about 40% – and the Portuguese education system, under the nationalist order, was mainly devoted to its internal borders and African colonies (MEC-OEI 2003). Again, recognition of the education gap in comparison with post–world war II European countries led the government to request assistance from the Organization for Economic Cooperation and Development (OECD) in 1955. The Mediterranean Regional Project, under the sponsorship of the OECD, marks the first attempt at alignment with the international education framework of the twentieth century (Alves 2012; Mendonça n.d.). Compulsory schooling was expanded to 4 years in the following year, although only for boys. Four years of mandatory schools for girls only arrived in 1960 (MEC-OEI 2003). Still, training of human resources was mostly ideological, valuing the nationalist ideology and the associated social promotion in cities, while the rural areas were lacking fully trained primary school teachers. The late 1950s saw an effort of the country's industrialization and the shift of rural populations to the big cities, exchanging farms for factories. The first report on the OECD's Mediterranean Project, released in 1964, pointed to the much-needed reform of the education system to answer the economy's dynamic requests (Alves 2012). In 1966, mandatory schooling for both sexes was extended to 6 years, and the first pedagogical research institute was created. Children who could not pursue their studies would do the 6 mandatory years, while those that were set to pursue further education, mainly determined by the socioeconomic status of their families, had to pass a national exam at the end of grade 4 before they could proceed to the lycées or technical education. Higher education was the logical end of the lycées, but only a small proportion of students at that time would pursue higher education. Professional/vocational training was the major concern of the state (MEC-OEI 2003).

At about the same time, the International Association for the Evaluation of Educational Achievement (IEA) was founded. In 1960, IEA deployed the first large-scale comparative educational assessment in 12 countries (the 'Pilot Twelve-Country Study'). During this and the following decades, IEA set much of the analytical framework for content assessment, student sampling, proficiency estimation and data analysis of international large-scale cross-national student assessments (ILSA). Since then, IEA has conducted ILSAs on mathematics, science and

reading literacy (e.g. FIMS, First International Mathematics Study in 1964 and FISS, First International Science Study in 1970–1971 preceded the 1995 TIMSS, Trends in International Math and Science Study) (IEA 2018).

Portugal – still under the 'new state' fascist regime – kept back facing the rest of Europe. It was only in the early 1970s that the minister of education Veiga Simão set the political context for the first large and profound reform of the Portuguese basic, secondary and higher education systems (MEC-OEI 2003). However, Simão's reform was never implemented since on 25 April 1974, a military coupe ended the 40-year-long fascist regime and set the country back to democracy. Despite the strong ideological and social conflicts that followed the revolution years, the importance of education on the country's social and economic development was consensual amongst all ideological parties and society strata (Alves 2012). Major changes led by the revolutionary spirit happened in those years. In 1975, the grade 4 exam was abolished, and students could not be retained anymore at grades 1 and 3. Mandatory schooling was extended to 9 years, and a major reorganization of schooling cycles and curricula occurred with the intent of increasing attendance of students, especially those from disadvantaged families, in primary education and lower secondary education (MEC-OEI 2003). The lycée and professional tracks were fused with a common core for grades 7 and 8, and students following the professional track were given access to higher education. Many professional schools were converted into university institutes which, like the classical universities, enjoyed renewed pedagogic, scientific and administrative autonomy. During the on-course revolution years, secondary education was completed with a civic year – where students worked to the benefit of their communities – preceding enrolment in higher education. In 1980 this civic year was replaced by grade 12 with the dual objective of being the terminal year for secondary education and the interim year for admission to higher education. Grade 12 had a dual pathway granting access to either the traditional 5-year sciences and humanities higher education degree or the professional oriented 3-year polytechnic higher education (MEC-OEI 2003). The harsh economic times of the 1980s and the need for skilled workers for the development of the fragile economy set the stage for basic and secondary education reform defined by 1986's 'Basis Law of the Educational System'. This law consigns that the right to education and culture for all children is 9 years of compulsory schooling, ensuring the training required for active participation in society, equality of opportunities, freedom of learning and teaching and the training of all young people and adults who had dropped school. Vocational and professional secondary courses that granted access to a profession or the pursuit of higher education were reformed in parallel with the regular sciences and humanities tracks. The basic and education system was organized in four cycles of study – the first cycle (grades 1–4), the second cycle (grades 5 and 6), the third cycle (grades 7–9) and secondary (grades 10–12). High-stake national exams were first introduced at grade 12 in 1994 with the dual purpose of certifying the end of the secondary education and rank students' access to higher education. National high-stake exams at the end of the third cycle of basic education (grade 9) for Portuguese language and mathematics were intro-

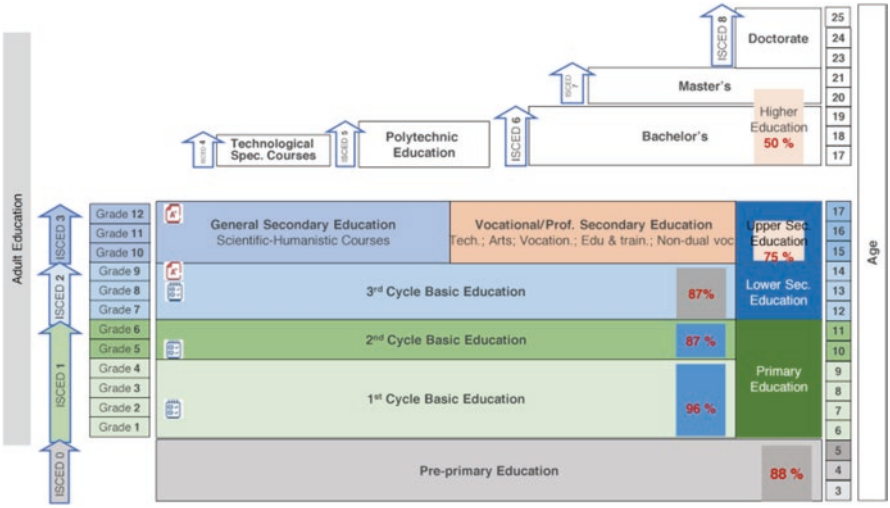

Fig. 17.1 The Portuguese education system (as of 2016). Percentage numbers are the coverage at each education cycle. and represent low-stakes and high-stakes national achievement test at the aligned grades. (Fundação Francisco Manuel dos Santos 2018; Marôco 2016; Ministério da Educação 2007; Organization for the Economic Cooperation and Development 2018)

duced in 2005, and national high-stake exams for Portuguese language and mathematics exams at grade 4 were introduced in 2013 (and revoked 3 years later with the change in government). In 2012, compulsory education was set to 12 years or until 18 years of age. Figure 17.1 summarizes the actual Portuguese education system.

Education policies and reforms in Portugal during the 1990s have been set by ideological agendas and interpretations of what were the best curricula, teaching and learning practices to reach the goals of social, cultural, scientific and economic development. However, little data-driven evidence was presented to support education policy and policy changes. At about the same time, OECD was feeling the need to fill the gap on valid and reliable, regularly collected, educational indicators to evaluate and compare the national education systems of its member countries. Thus, the Programme for International Student Assessment (PISA) was created in 2000 to respond to what OECD claimed to be the lack of quality and coverage of cross-national student achievement data generated by IEA studies started in the 1960s (Breakspear 2014).

Aimed at gathering external data on the Portuguese students' knowledge and skills, Portugal participated in the first edition of TIMSS (1995) but then withdraw from the study until 2011. The next participation of Portugal on an ILSA was in the first cycle of PISA, in 2000. Ever since, the evolution and influence of these international large-scale students' assessments on the educational policies support and need for reforms have been quite substantive as we shall see in the next sections.

International Assessments Today

Following the very poor results obtained by Portuguese students in the 1995 TIMSS edition with grade 4 and grade 8 students, Portugal withdraw from this ILSA because decision makers at that time felt that TIMSS was not a valid measure of the Portuguese students' specific knowledge and skills which were not aligned with the TIMSS curricula framework. It was only in 2011 that Portugal returned to TIMSS and PIRLS (Progress in International Reading Literacy study) for grade 4, participating again in the 2015 TIMSS (grade 4) and TIMSS Advanced (grade 12) editions and in the 2016 PIRLS and ePIRLS (electronic PIRLS). The ninth grade Portuguese students' proficiency with foreign languages (English and French) was assessed in 2011 with the First European Survey on Language Competence (*SurveyLang*) project sponsored by the European Commission. Being an OECD member, Portugal participated in the first PISA cycle (2000) and all other cycles that followed (2003, 2006, 2009, 2015 and 2018). In 2018, Portugal also participated for the first time in the IEA's International Computer Information Literacy Study (ICILS 2018). In 2019, Portugal is set to participate in TIMSS for grade 4 and TIMSS for grade 8. For 2021, participation in the PISA 2021 and PIRLS 2021 is planned, and preparation for these studies has already started. It is worth to note that PISA 2015 and subsequent cycles were done as computer-based assessment. Portugal was also one of the 21 countries that did, in 2017, the computer-based eTIMSS pilot study. TIMSS 2019 and PIRLS 2021 are planned to be delivered in a computer-based (e-assessment) format.

Nowadays, ILSA data is seen by education policy makers as fundamental for the external evaluation of the education system, to benchmark the evolution of the basic and secondary students' knowledge and skills and to support education policies (for a review, see Afonso and Costa 2009; Carvalho et al. 2017). The impact of ILSA results on the education community, policy makers and the public is well illustrated by the media that profusely report the ILSA results obtained by the Portuguese students and commentaries on the results, causes and consequences by education specialists and policy makers alike. In the days and week after the TIMSS 2015 results release (29/11/2016) more than a dozen text, radio and TV reports were published or aired. The same metric was observed for PISA 2015 results (released on 06/12/2016) that were extensively reported by the media in the days and week after the release. The education community, both from teacher training institutes and universities as well as the private sector, has devoted much attention to the ILSA results and secondary data analysis. For example, the Fundação Francisco Manuel dos Santos, a private philanthropic foundation, has sponsored and published secondary analysis of the Portuguese PISA results targeted to the public and educators (Ferreira et al. 2017a, b). The National Education Council, a policy consulting agency to the Ministry of Education, has set a programme of conferences and publications on the ILSA secondary analysis, again targeted mainly for educators and the public (Conselho Nacional de Educação 2013, 2015). However, peer-reviewed research on Portuguese ILSA data published in specialized journals is still scarce. A

recent review by Carvalho et al. (2017) regarding PISA secondary analysis only identified nine research papers, published mainly in economy and management journals, from the 2009–2015 PISA period.

Portugal Performance on the ILSAs

ILSA results in Portugal have been somewhat contradictory. While there was a consistent trend in all three domains of PISA, a feature that had no parallel in the European Union, and in TIMSS mathematics, the same pattern was not observed for TIMSS science and PIRLS reading literacy. Figure 17.2 illustrates the Portuguese results in PISA, TIMSS fourth grade and PIRLS from 1995 to 2016. From the bottom of the table of the OECD countries in the PISA 2000 cycle, Portuguese students have jumped approximately one half of a standard deviation on the PISA scale (that is almost two years of formal schooling) in 15 years. The average growth rate was 1.8 PISA points per year for reading literacy, 2.6 PISA points per year for mathematical literacy and 2.8 PISA points per year for scientific literacy. This is particularly relevant since OECD countries overall, in the same period, showed a negative growth rate of −0.6 PISA points per year for reading literacy, −0.5 PISA points per year for mathematical literacy and −0.3 PISA points per year for scientific literacy (Fig. 17.2.). In the 2015 cycle, Portuguese students ranked significantly above the OECD average for scientific and reading literacy, being on the OECD average for mathematical literacy (Marôco et al. 2016a). For TIMSS and PIRLS, there are only

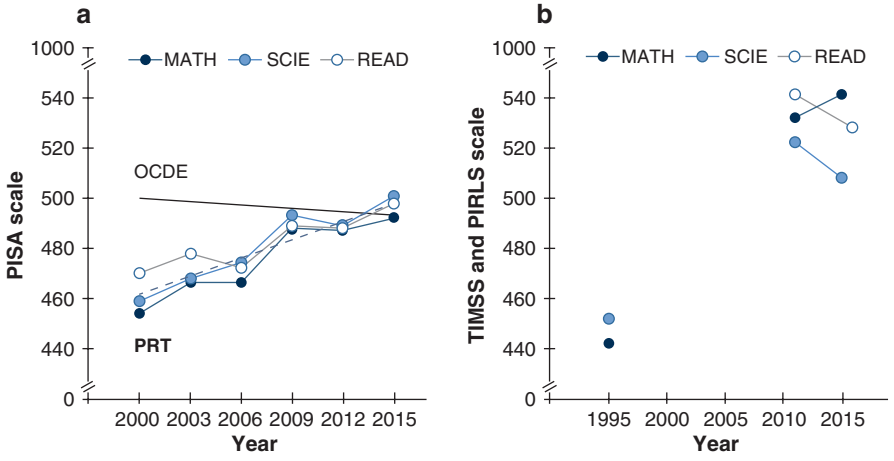

Fig. 17.2 PISA, TIMSS and PIRLS mean score for Portugal from 2000 to 2015 (PISA) and 1995 to 2015–2016 (TIMSS) and 2011 to 2016 (PIRLS). In panel A (PISA), the broken line represents the Portuguese evolution in all three PISA domains (2.4 points/year), while the continuous line represents the OECD mean evolution in the three domains (−0.5 points/year). (Adapted from Marôco et al. 2016a, b)

data for three and two of its cycles, respectively. There was a quite relevant evolution from TIMSS 1995 to TIMSS 2011. However, for 2015 the trend is not clear. While for mathematics at grade 4, the Portuguese students ranked significantly above the international mean score and even got better results than those from countries who are generally pointed to for reference in the ILSA constellation (e.g. Finland), there were statistically significant drops for science (TIMSS) and reading literacy (PIRLS) from the penultimate to the last editions of these studies (Marôco 2018; Marôco et al. 2016b). TIMSS Advanced (for grade 12) results positioned Portugal in the middle of the ranked table of participants, significantly above all the European countries that took part in the TIMSS Advanced mathematics and physics test (Marôco et al. 2016c).

Despite the praised positive evolution, even by OECD's Andreas Schleicher who stated that 'Portugal is Europe's biggest success story in PISA' (Tavares 2017), PISA does reveal some of the fragilities of the Portuguese education system. One of the most striking is the strong regional asymmetries in the PISA results. Early secondary analysis by Pereira and Reis (2012) with PISA 2009 mathematics and reading literacy data revealed statistically significant regional differences with the autonomous region of Madeira, south and interior regions scoring 30–40 points below the national average. Those differences were maintained or even amplified in PISA 2015 (Marôco 2017). The statistically significant lower performing NUTS III regions are still located in the interior, north and autonomous archipelagos regions. The top-performing regions (with average scores significantly above the national mean by more than 10 points) are generally located on the coastal and more developed regions. This pattern, as illustrated by Fig. 17.3 for mathematical literacy results in the 2015 TIMSS and PISA editions, is similar for scientific and reading literacies (data not shown).

Further research, with PISA 2015 scientific literacy data using with hierarchical linear modelling with regions as clusters, revealed that the epistemic beliefs in

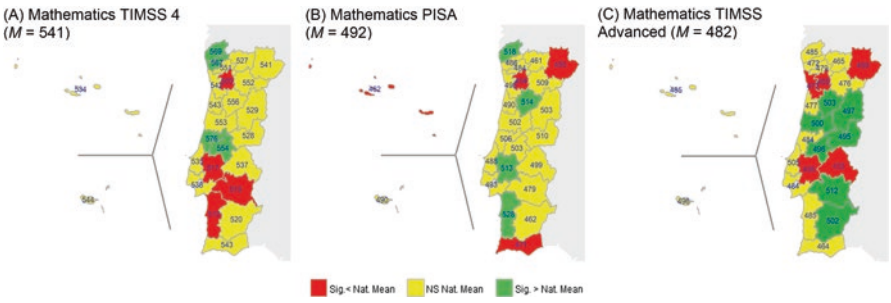

Fig. 17.3 Regional scores for mathematical literacy at (**a**) TIMSS 4 2015 [grade 4], (**b**) PISA 2015 [grades 7–11, with 20% and 57% of PISA students at grades 9 and 10, respectively] and (**c**) TIMSS Advanced 2015 [grade 12]. Dark red indicates regions with mean scores statistically significantly below the national average, yellow indicates regions with mean scores not statistically different from the national mean (NS), and green indicates regions with mean scores statistically significantly above the national average (M) for $p < 0.05$. (Marôco et al. 2016a, b, c)

science ($\beta = 0.27$, $p < 0.05$), the expected occupational status by age 30 (BSMJ; $\beta = 0.26$, $p < 0.05$) and the economic and sociocultural status (ESCS; $\beta = 0.23$, $p < 0.05$) at the student level and the students behaviours hindering learning ($\beta = -0.31$, $p < 0.05$) and class size ($\beta = 0.39$, $p < 0.05$) at the school level were the strongest predictors of the observed regional scores' variation in PISA 2015 (Marôco 2017). A similar analysis for PIRLS 2016 results revealed that the best predictors for reading literacy of Portuguese fourth graders were the confidence as readers ($\beta = 0.31$, $p < 0.05$) and the home resources for learning ($\beta = 0.26$, $p < 0.05$) at the student level. The emphasis of schools on academic success ($\beta = 0.40$, $p < 0.05$) and the desire of students to do well in school ($\beta = 0.31$, $p < 0.05$) were the best predictors of reading literacy at the school and teacher levels (Marôco 2018).

Concerns regarding the validity of ILSAs to assess the Portuguese students' knowledge and skills have been verbalized both by policy makers and the media (Cristo 2017). ILSAs, like PISA or TIMSS and PIRLS, are generally presented as reliable transnational instruments to benchmark the performance of education systems and facilitate education reforms (Breakspear 2012; Carvalho 2012; Nelson 2002; Phillips and Jiang 2015). However, criticisms have been drawn on these ILSA uses due to the lack of transcultural invariance (Rutkowski and Svetina 2014) and lack of evidence of the external validity of the ILSA as regarded to their driven education policies change (Hanberger 2014) and reliability and criterion validity (Schult and Sparfeldt 2016). At the present date, to the best of my knowledge, there is but one available report on the concurrent validity of ILSA estimated literacies and countries' national assessments: APavešić and Cankar (2018) observed a 0.7 correlation between TIMSS Advanced mathematics and the mathematics natura exam in Slovenia in 2015. Although national assessments, like high-stake certification and graduation exams, have different objectives from the low-stake ILSA, assessed domains are, in different degrees, shared, and thus concurrent validity should be observed. Using the mathematical literacy as an example, we conducted a national vs. ILSA tests content and correlation analysis of the national high-stake mathematics exams scores with mathematical literacies evaluated by TIMSS at grades 4 and 12 and PISA at grades 9 and 10 (Marôco and Lourenço 2017). Table 17.1 summarizes the common features of the 2015 Portuguese national high-stake exams and the TIMSS fourth grade, PISA and TIMSS Advanced content and cognitive domains. The content domains of the national high-stake exams are better aligned with the TIMSS fourth grade and TIMSS Advanced and somewhat less with PISA. This is easily explained by the class/curriculum-based TIMSS as compared to the age-based PISA, as well as the policy changes to better align the national curricula with the TIMSS frameworks. Analysis of the national high-stake exams results of students that participated on the 2015 cycles of TIMSS and PISA found moderate to strong correlations between the scores of ILSA's mathematical literacies and the national mathematics exams (Fig. 17.4). The observed correlations were higher for TIMSS fourth grade and TIMSS Advanced ($r = 0.71 \pm 0.01$, $p < 0.001$) than for PISA ($r = 0.63 \pm 0.01$, $p < 0.001$). It is worthwhile to mention that the magnitudes of these correlations were similar to the correlations between

Table 17.1 Content analysis of the Portuguese mathematics national exams for grades 4, 9 and 12 during the 2014/2015 school year (first phase) and the mathematics literacy tests of TIMSS grade 4, PISA and TIMSS Advanced (2015 editions) (Marôco and Lourenço 2017)

Domains	ILSA mathematics (%)	National mathematics exam (%)
	2015 TIMSS grade 4	Grade 4 (2014/2015)
Content domains		
Numbers	50	44
Geometric shapes and measures	35	43
Data display	15	13
Cognitive domains		
Knowing	40	43
Applying	40	36
Reasoning	20	21
	2015 PISA	Grade 9 (2014/2015)
Content domains		
Quantity	25	12
Space and shape	25	40
Change and relationships	25	35
Uncertainty and data	25	13
Cognitive domains		
Knowing	–	37
Formulating	25	–
Applying	50	38
Interpreting/reasoning	25	2
	2015 TIMSS Advanced	Grade 12 (2014/2015)
Content domains		
Algebra	35	20
Calculus	35	32
Geometry	29	27
Probability and combinatorics	–	20
Cognitive domains		
Knowing	29	23
Applying	41	57
Reasoning	30	20

the sampled students' final mathematics score (assigned by teachers) and their score on the national exams (r = 0.68 for grade 4; r = 0.62 for grade 9; r = 0.77 for grade 12).

It is noticeable that despite the economic crisis of the 2008–2013 period when national GDP was reduced by 8% (Perez and Matsaganis 2018), Portugal was still able to increase its overall PISA score, ranking in the 2015 edition, for the first time, above the OECD average for scientific, reading literacies. Ferreira et al. (2017a) did a comparative study of the PISA results from 2000 to 2015 and identified, as follows, the principal positive and negative features that explain Portugal's evolution in PISA. Although Portugal is a relatively poor country as compared to

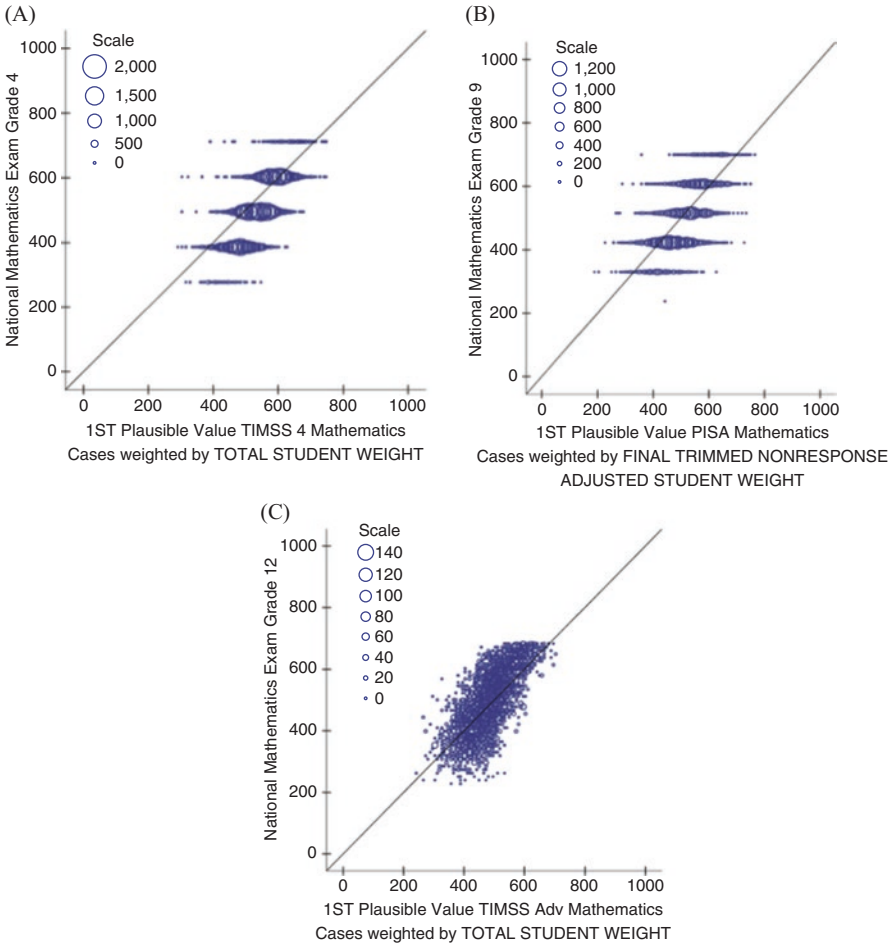

Fig. 17.4 Concurrent validity of the national mathematics exams at grades 4, 9 and 12 and the TIMSS grade 4 (**a**), PISA (**b**) and TIMSS Advanced (**c**) mathematics literacies (first plausible value). National exam scores for grades 4 and 9 range from 1 to 5. Scores for grade 12 range from 0 to 20. National exam scores were converted to the TIMSS and PISA scale (M = 500, SD = 100) before correlation analysis. (Marôco and Lourenço 2017)

other OECD members (the 2015 GDP per capita for Portugal was 29.5 kUSD vs. 41.1 kUSD for the OECD mean (OECD.Stat n.d.), the expenditure in education per capita is in line with other OECD member states. Pre-school coverage is close to 100%; teachers have appropriate specific and pedagogical training and feel they can make a change in students' lives. Students are motivated and persistent and feel supported by their parents and teachers. Thirty percent of schools in less favoured economic regions have better results than what would be expected by their ESCS. These results were attributed to teachers' competence, motivation and openness towards change, students' engagement and discipline and schools'

educational projects aligned with the community. However, there is still much need for improvement in the Portuguese education system, namely, on the reduction of grade retention rates, increase in parental education, renewal of aging teachers and improvement in schools' autonomy, especially as far as teacher recruitment is concerned (Ferreira et al. 2017a).

Impacts of ILSAs on the Portuguese Education System

TIMSS 1995 showed that Portuguese students were significantly behind the students in the Western countries that took the test. According to the education policy makers in office during that period (1995–2000), these poor results were attributed to the maladjustment of the curricula and typology of the TIMSS test to the Portuguese education system (Barroso 2010). This undisputed perception of the negative TIMSS results and its validity for the national context were used as a pretext to delay the discussion on the national curricula, to analyse the TIMSS results and to promote their dissemination and discussion (Barroso 2010; Carvalho et al. 2017). However, and despite the devaluation of TIMSS, the Institute for Educational Innovation, which was responsible for the TIMSS 1995 application in Portugal, started in 1996 a national, sample-based, diagnostic test of mathematics whose conceptual inspiration came clearly from the TIMSS 1995 framework (Amaro 1997; Barroso 2010). Results from TIMSS 1995 raised the awareness of the need for the external assessment of students. This was explicitly recognized with the creation, in 1997, of the Office for Educational Evaluation (GAVE), a central office of the Ministry of Education, whose responsibilities were for the planning, coordination, creation and validation of external learning assessment instruments, as well as the coordination of future participation in ILSAs (Justino and de Almeida 2017). In 2013, the GAVE responsibilities were passed on to the Institute for Educational Evaluation (IAVE), a public institute with scientific and administrative autonomy under the supervision of the Ministry of Education. The influence of the TIMSS and PISA frameworks on the national basic and secondary exams as well as the importance of the external assessments, both low and high stakes, driven by earlier ILSA participation, still prevails in the Portuguese education system.

Other than the gains in methodological and external assessment practices, no real or significant perceived educational policy consequences were driven from the Portuguese participation in the ILSA during the 1991–2003 period (Fernandes 2014; Fernandes and Gonçalves 2018). It is in the first decade of the twenty-first century, specially from 2005 on, that the education policy discourse included explicit references to PISA results (Carvalho et al. 2017). Before this date, PISA reports have been evoked only twice: to support the reorganization of the basic and secondary education curricula in 2001 by the education minister Julio Pedrosa following the PISA 2000 results and, in 2004, during a Parliament session, when the minister of education Carmo Seabra quotes the 2003 PISA results to prioritize the

learning of the Portuguese language, mathematics and science in the national curricula (Afonso and Costa 2009). The next education minister, Maria de Lurdes Rodrigues (2005–2009), explicitly used the PISA results, in line with the national low-stakes test results, to promote data-driven policy measures (Afonso and Costa 2009; Carvalho et al. 2017; Fernandes and Gonçalves 2018). The first National Programme for Teaching the Portuguese Language, the National Reading Plan and the Mathematics Action Plan were justified on the basis of the much needed improvement of Portuguese students in the PISA tests, despite the fact that these plans were in line with the educational policies started in the early 1990s (Fernandes 2014; Fernandes and Gonçalves 2018). The reforms made by curricular changes in the first decade of the twenty-first century, inspired by TIMSS and PISA, were accompanied by a strengthening of external assessment of students and schools, namely, with the high-stakes standardized exams for mathematics and Portuguese language at grade 9 promoted by the 2002–2004 minister of education David Justino (Justino and de Almeida 2017). PISA data was also used to support policies aimed at the enlargement of economic support to students from low-income families, facilitate the access to internet and computers for primary education (the 2007 Technological Education Plan), and to reorganize the Priority Intervention Educational Territories (TEIP) Program for schools located in economic depressed areas (Afonso and Costa 2009). Furthermore, under the expertise of the PISA reports on teachers qualifications and professional practices, controversial policies like the teachers' performance assessment and the revision of the qualifications required to become a teacher were proposed during the 2005–2011 period (Afonso and Costa 2009; Carvalho et al. 2017) although some, like the Teachers' Exam of Knowledge and Capacities (PACC), required to land a teaching job, were only later (2014–2015) and, briefly, implemented.

ILSA results and frameworks, namely, PISA and TIMSS, were again recalled in the 2011–2015 period. The education minister Nuno Crato serving during this term explicitly quoted PISA and TIMSS required skills and competencies to further reform and strengthen the mathematics, sciences and Portuguese language curricula and targets, to recommend teaching practices and timetables and to further expand the national high-stakes exams portfolio to grades 4 and 6. The importance of the alignment of the national curricula to the ILSA curricula was recognized in the established national advanced mathematics curriculum that stated: 'analysis of these elements [TIMSS Advanced framework], as well as curricula from other countries not participating in TIMSS Advanced, reveals that the inclusion in the curriculum of some fundamental themes, currently absent from the Secondary Education in Portugal, contributes decisively to the alignment of national curricular options with the international plan (…)' (Bivar et al. 2015). Figure 17.5 summarizes the major ILSA policy drivers in the Portuguese education system aligned with the PISA mathematics literacy score and the national education expenditure.

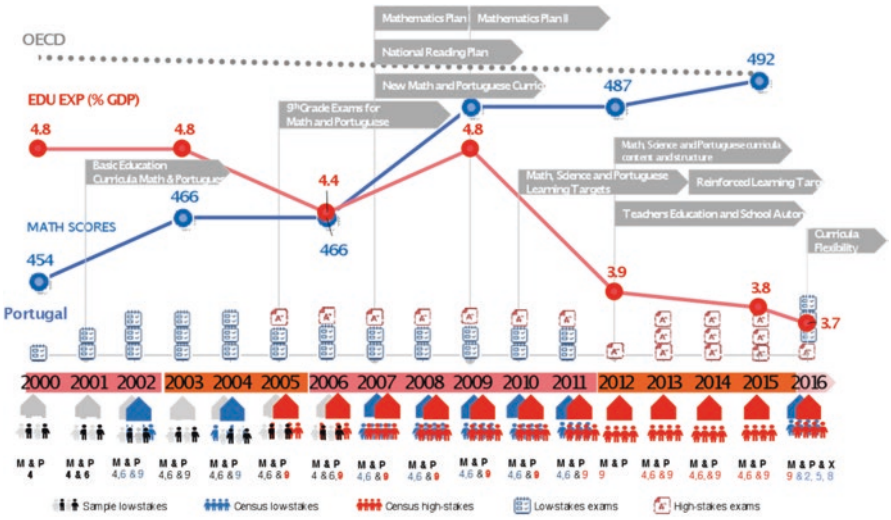

Fig. 17.5 Major ILSA-*driven* education policy and external assessments changes during the PISA timeframe. Pink bars represent socialist governments, orange bars represent social democrats' governments. M stands for mathematics, P stands for Portuguese language, X stands for multiple subjects (geography, history, etc.) at different grades (4, 6, 9, etc.) and EDU EXP stands for educational expenditure, as percentage of the GDP in the PISA years. (Adapted from Ministério da Educação e Ciência 2015 and references in the text)

Concluding Remarks

My analysis shows that ILSA results in Portugal give valid and reliable indicators of the Portuguese national education system performance per comparison to international frameworks. Trends for Portuguese students' performance can be inferred from participation in ILSA. This feature of ILSA compensates for the limitation of the Portuguese national exams system, which, being public, doesn't have trend items that could be used to build trends. ILSA has also been used to build knowledge on both large-scale assessments and best practices for standardized testing. The Portuguese education reforms were, from the early years of the twentieth century to the present day, driven by the recognition of the system's deficiencies as compared with other European countries. International diagnostics, done by organizations like the OECD, and results from early participation in ILSA in the late twentieth century have set the stage for educational policy change and to support system-wide interventions, reshaping the Portuguese education landscape in the twenty-first century.

References

Afonso, N., & Costa, E. (2009). The influence of the Programme for International Student Assessment (PISA) on policy decision in Portugal: the education policies of the 17th Portuguese Constitutional Government. *Sísifo. Educational Sciences Journal, 10*(set/dez), 53–64. Retrieved from http://sisifo.fpce.ul.pt

Alves, L. A. M. (2012). *História da Educação – Uma Introdução*. Porto. Retrieved from http://ler.letras.up.pt/uploads/ficheiros/10021.pdf

Amaro, G. (1997). Qualidade em Educação: a avaliação externa das aprendizagens dos alunos em Portugal. *Inovação, 10*(2–3), 259–275.

Barroso, C. F. de A. (2010). *Os estudos PISA e o ensino das ciências físico-naturais em Portugal: a comparabilidade dos resultados nacionais e as implicações para a política educacional*. Universidade Nova de Lisboa. Retrieved from https://run.unl.pt/bitstream/10362/5279/1/carlosbarroso.pdf

Bivar, A., Grosso, C., Oliveira, F., & Timóteo, M. (2015). *Programa e Metas Curriculares. Matemática. Ministério da Educação e da Ciência*. Lisboa. Retrieved from http://www.dge.mec.pt/sites/default/files/ficheiros/programa_metas_curriculares_matematica_a_secundario.pdf

Breakspear, S. (2012). The policy impact of PISA: An Exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Papers, 71*, 1–31. https://doi.org/10.1787/5k9fdfqffr28-en.

Breakspear, S. (2014). How does PISA shape education policy making? Why how we measure learning determines what counts in education. *Centre for Strategic Education: Seminar Series, 240*, 16. Retrieved from http://simonbreakspear.com/wp-content/uploads/2015/09/Breakspear-PISA-Paper.pdf.

Candeias, A., Paz, A., & Rocha, M. (2007). *Alfabetização e escola em Portugal nos séculos XIX e XX: os censos e as estatísticas*. Lisboa: Fundação Calouste Gulbenkian.

Carvalho, L. M. (2012). The fabrications and travels of a knowledge-policy instrument. *European Educational Research Journal, 11*(2), 172–188. https://doi.org/10.2304/eerj.2012.11.2.172.

Carvalho, L. M., Costa, E., & Gonçalves, C. (2017). Fifteen years looking at the mirror: On the presence of PISA in education policy processes (Portugal, 2000–2016). *European Journal of Education, 52*(2), 154–166. https://doi.org/10.1111/ejed.12210.

Conselho Nacional de Educação. (2013). *Avaliações internacionais e desempenho dos alunos portugueses*. (A. M. Bettencourt & M. Miguéns, Eds.) (Seminários). Lisboa: Conselho Nacional de Educação.

Conselho Nacional de Educação. (2015). *Investigação em Educação e os Resultados do PISA*. (A. L. Ferreira, P. Félix, & S. Ferreira, Eds.) (Seminários). Lisboa: Conselho Nacional de Educação. Retrieved from http://www.cnedu.pt/content/edicoes/seminarios_e_coloquios/PISA_Investigação_em_Portugal_dezembro_2014.pdf

Cristo, A. H. (2017). Os três problemas dos exames nacionais. *Observador*. Retrieved from https://observador.pt/especiais/os-tres-problemas-dos-exames-nacionais/

Fernandes, D. (2014). Avaliações externas e melhoria das aprendizagens dos alunos. In CNE (Ed.), *Avaliação Externa e Qualidade das Aprendizagens* (pp. 21–49). CNE: Lisboa. Retrieved from http://www.cnedu.pt/content/edicoes/seminarios_e_coloquios/Avaliacão_externa_e_qualidade_das_aprendizagens_vf.pdf.

Fernandes, D., & Gonçalves, C. (2018). Para Compreender O Desempenho Dos Alunos Portugueses No PISA (2000-2015). In M. I. R. Ortigão (Ed.), *Políticas de Avaliação, Currículo e Qualidade: Diálogos sobre o PISA – Volume 3* (pp. 39–68). Curitiba: EDITORA CRV. https://doi.org/10.24824/978854442369.1.

Ferreira, A. S., Flores, I., & Casas-Novas, T. (2017a). *Porque Melhoraram os Resultados PISA em Portugal. Estudo Longitudinal e comparado (2000–2015)*. Retrieved from https://www.ffms.pt/FileDownload/9857244f-4dfb-48ad-b196-0448dc444865/porque-melhoraram-os-resultados-pisa-em-portugal

Ferreira, A. S., Simões, B., Flores, I., Leiria, I., & Casas-Novas, T. (2017b). *A educação em Exame.pt* [Education under Exam.pt]. Retrieved July 10, 2018, from https://educacaoemexame.pt

Fundação Francisco Manuel dos Santos. (2018). *PORDATA – Taxa bruta de escolarização por nível de ensino*. Retrieved July 10, 2018, from https://www.pordata.pt/Portugal/Taxa+bruta+de+escolarização+por+nível+de+ensino-434-7644.

Hanberger, A. (2014). What PISA intends to and can possibly achieve: A critical programme theory analysis. *European Educational Research Journal, 13*(2), 167–180. https://doi.org/10.2304/eerj.2014.13.2.167.

IEA. (2018). *Brief History of the IEA*. Retrieved July 5, 2018, from https://www.iea.nl/brief-history-iea

Justino, D., & de Almeida, S. (2017). International assessment, curriculum policy induction and instruction time management: Lessons from Portuguese experience. *European Journal of Curriculum Studies, 4*(2), 671–691. Retrieved from http://pages.ie.uminho.pt/ejcs/index.php/ejcs/article/view/162

Marôco, J. (2016). *Portugal – TIMSS 2015 Encyclopedia*. Retrieved September 21, 2018, from http://timssandpirls.bc.edu/timss2015/encyclopedia/countries/portugal/

Marôco, J. (2017). Assimetrias Educacionais em Portugal: Através das Lentes do PISA. In Conselho Nacional de Educação (CNE) (Ed.), *Estado da Educação 2016* (pp. 254–274). Lisboa: Conselho Nacional de Educação (CNE). Retrieved from http://www.cnedu.pt/content/noticias/CNE/CNE-EE2016_web.pdf

Marôco, J. (2018). O Bom Leitor: Preditores da Literacia de Leitura dos Alunos Portugueses no PIRLS 2016. *Revista Portuguesa de Educação, 31*(2), 115-131. Retrieved from https://revistas.rcaap.pt/rpe/article/view/13768

Marôco, J., Gonçalves, C., Lourenço, V., & Mendes, R. (2016a). *PISA 2015 – Portugal: Literacia científica, literacia de Leitura & literacia matemática* (Vol. I). Lisbon: IAVE. Retrieved from http://iave.pt/images/FicheirosPDF/Estudos_Internacionais/Relatorio_PISA2015.pdf.

Marôco, J., Gonçalves, C., Lourenço, V., & Mendes, R. (2016b). *TIMSS 2015 – Portugal. Volume I: Desempenhos em Matemática e em Ciências*. Retrieved from http://iave.pt/images/FicheirosPDF/Estudos_Internacionais/TIMSS/Relat_rio_TIMSS4.pdf

Marôco, J., Gonçalves, C., Lourenço, V., & Mendes, R. (2016c). *TIMSS advanced 2015 - Portugal: Desempenhos em matemática e em ciências* (Vol. 1). Lisbon: IAVE. Retrieved from http://iave.pt/images/FicheirosPDF/Estudos_Internacionais/TIMSS/Relatorio_TA2015.pdf

Marôco, J., & Lourenço, V. (2017). In Search of Concurrent Validity of Large-scale International Students Assessments (LSISA) and High-stakes National Exams: The Portuguese Case Study. In *Paper presented at the 2017 European Conference on Educational Research: Network 9 Assessment, Evaluation, Testing and Measurement* (p. 126). Copenhagen. Retrieved from http://www.iave.pt/images/FicheirosPDF/Estudos_Internacionais/ECER2017_InSearchOfConcurrentValidity.pdf

MEC-OEI. (2003). *Breve Evolução Histórica Do Sistema Educativo*. Lisboa. Retrieved from https://www.oei.es/historico/quipu/portugal/historia.pdf

Mendonça, A. (n.d.). *Evolução da Politica Educativa em Portugal*. Retrieved from http://www3.uma.pt/alicemendonca/conteudo/investigacao/evolucaodapoliticaeducativaemPortugal.pdf

Ministério da Educação. (2007). *Educação e Formação em Portugal* (M. da Educação, Ed.). Lisboa: Editorial do Ministério da Educação. Retrieved from http://www.dgeec.mec.pt/np4/97/%7B$clientServletPath%7D/?newsId=147&fileName=educacao_formacao_portugal.pdf

Ministério da Educação e Ciência. (2015). *Educação Pré-Escolar, Ensino Básico e Secundário – Relações Internacionais 2011–2015* (Ministério da Educação e Ciência, Ed.). Lisboa.

Nelson, D. I. (2002). *CPRE policy briefs: Using TIMSS To inform policy and practice at the local level.* Philadelphia. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED477658&lang=pt-br&site=ehost-live&scope=site

OECD.Stat. (n.d.). *Level of GDP per capita and productivity*. Retrieved July 13, 2018, from https://stats.oecd.org/Index.aspx?DataSetCode=PDB_LV

Organization for the Economic Cooperation and Development. (2018). *Education GPS – Portugal*. Retrieved July 10, 2018, from http://gpseducation.oecd.org/CountryProfile?primaryCountry=PRT

Pereira, M. C., & Reis, H. (2012). Diferenças regionais no desempenho dos alunos portugueses: Evidencia do programa PISA da OCDE. *Boletim Económico Banco Portugal, 3*(Inverno), 59–83. Retrieved from https://www.bportugal.pt/paper/diferencas-regionais-no-desempenho-dos-alunos-portugueses-evidencia-do-programa-pisa-da-ocde

Perez, S. A., & Matsaganis, M. (2018). The political economy of austerity in Southern Europe. *New Political Economy, 23*(2), 192–207. https://doi.org/10.1080/13563467.2017.1370445.

Phillips, G. W., & Jiang, T. (2015). Using PISA as an international benchmark in standard setting. *Journal of Applied Measurement, 16*(2), 161–170.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57. https://doi.org/10.1177/0013164413498257.

Schult, J., & Sparfeldt, J. R. (2016). Reliability and validity of PIRLS and TIMSS. Does response format matter? *European Journal of Psychological Assessment*, 1–12. https://doi.org/10.1027/1015-5759/a000338

Tavares, P. S. (2017). Andreas Schleicher: "Portugal é a maior história de sucesso da Europa no PISA". *Diário de Notícias*. Retrieved from https://www.dn.pt/portugal/interior/andreas-schleicher-portugal-e-a-maior-historia-de-sucesso-da-europa-no-pisa-5659076.html

Chapter 18
International Assessment Studies in Serbia Between Traditional Solutions, Unexpected Achievements and High Expectations

Dragica Pavlović Babić

Since 2001, Serbia has been participating in IEA/TIMSS (grade 4, starting from 2011) and OECD/PISA assessment studies. During this period, assessment context was characterized by the highly centralized and over-controlled system with content-based curricula and traditional teaching methods which set students in a passive position with general expectations placed on the level of literate reproduction of the poorly integrated facts, underdeveloped assessment system and lack of the assessment data. International assessment studies revealed that achievements were disappointingly low and statistically below the international average in all examined domains with a high percentage of students below the level of functional literacy and a very small number of them on the highest proficiency levels. But, at the same time, the equity on the OECD average was high; the younger students (grade 4) performed better than their peers, with the achievement over the international average; literacy progress of about one half of the standard deviation in reading is made, which is further discussed in detail. So far, assessment data have influenced some legal and strategic solutions, but have not made any visible influence on the curricula, teaching methodology, assessment practice in school and in-service teacher education.

The History and the Context of International Assessment in Serbia

The history of international assessment in Serbia relates to the history of the International Association for the Evaluation of Educational Achievement (IEA), which organized and operated in cooperation with the University of Chicago under

D. Pavlović Babić (✉)
University of Belgrade, Belgrade, Serbia
e-mail: dragica.pavlovic@f.bg.ac.rs

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_18

the leadership of Torsten Hussen from 1959 to 1962, a small-scale study in 12 countries. One of those countries was Yugoslavia, and Serbia as one of its constitutive republics. In that study, about 10,000 children aged 13–14 years were tested in reading comprehension, mathematics, science, geography and non-verbal ability (de Landsheere 1974).

From this early participation, Serbia faced serious political changes, including two state breakups (in 1992 and 2006) as a result of a series of political upheavals, conflicts and wars during the 1990s. In that almost 30 years lasting period, the educational system remained closed to transformation processes and was increasingly isolated from the international community (Dimou 2009), partially due to the international sanctions which banned, among others, scientific, cultural, educational and technical exchanges. The sanctions ended in October 2000, after major internal political changes.

Immediately after the following year, Serbia joined two international assessment studies: Trends in International Mathematics and Science Studies, grade 8 (TIMSS) organized by the IEA, and Programme for International Student Assessment (PISA) by the Organization for Economic Co-operation and Development (OECD). Both studies were accomplished for the first time in 2003.

Highlights of the Educational System in Serbia

The main structure of the educational system in Serbia remains almost unchanged for decades and it consists of compulsory primary education divided into two four-year cycles for 6½–14½-year-old students; secondary education which consists of 4 years of general education or alternatively 2–4 years of vocational education for students older than 14½ years; and tertiary education. Since 2004, preparatory preschool programmes lasting for a half of a school year, were introduced as obligatory (Vujisić-Živković 2015).

At all pre-university education levels a high coverage and continuing increase of coverage was identified (Ministry of Education, Science and Technological Development of Republic of Serbia 2018). The coverage rate of primary education in Serbia is very high (98%), although the number of expert analyses stated that the actual rate is somewhat lower due to dropout rates that is found to be higher than those officially published (Šolić-Vojinović and Nastić-Stojanović 2014).

Lack of the Assessment Learning System

The current state in the area of assessment previously could hardly be recognized as the system, but some of the elements, although few and irregularly implemented, exist and have a certain impact on educational policies. Some of them include:

National Assessment Studies Before introducing international assessment studies PISA and TIMSS, the only external assessment findings were through two national assessment studies implemented in the 1990s (Havelka et al. 1990; UNICEF 2001), both conceptualized and realized by the independent research institution. Both of them revealed that students at the end of primary school display rather poor results in all examined domains.

In and after the 2000s, national assessment studies were accomplished at the end of the first cycle of the primary education (ZVKOV 2005, 2007), with significant impact on the process of defining national educational standards for this educational level.

National Examinations At the end of primary education, all students are obliged to attend the final examination. Since 2011, the final examination with the same functions is based on the previously adopted educational standards. The examination is conducted externally, while the final examination at the end of secondary education is traditionally administered at the school level simultaneously. But, this examination is currently in the transformation process, and it is expected that new concept will be implemented for the first time at the end of the 2020/2021 school year. The new "matura" examination will be external and it is planned to serve as qualification and entrance examination for the tertiary education (MPNS 2017).

School Marking System The only data which the system has used permanently for the assessment of learning achievement are school marks. But, school marking is not standardized, but rather relies on teachers' individual benchmarks and beliefs, a desirable outcome at the student level (UNICEF 2001). Educational statistics show that the average grades of students are constantly increasing (ZVKOV 2017). Thus, school marks over the years has become less informative for educational policy.

Snapshot of Findings of the International Assessment Studies

TIMSS 2003 and 2007, as well as all PISA cycles (2003, 2006, 2009, 2012) have showed that students from Serbia are behind the international average by about 1.5 schooling year, or almost half of a standard deviation (S.D.). For example, the average performance of students in Serbia in mathematics in 2012 (Mean: 449, S.E: 3.4; S.D.: 91 and S.E: 2.2) is for 45 lower than the OECD average (Mean: 494, S.E: 0.5; S.D.: 92 and S.E: 0.3).

Serbia also has a much larger share of students that do not achieve level of functional literacy. The EU 2020 target is to have not more than 15% of students below PISA proficiency level 2. In Serbia, according to PISA 2012, almost 40% of students are considered functionally innumerate (below level 2), while 33% of students are functionally illiterate in reading, and 35% functionally illiterate in science (OECD 2014b).

Over time across PISA assessments, comparing 2006 and 2012, students from Serbia realized the statistically significant positive annual change of 2.2 (S.E. 0.93). They also perform in problem solving (Mean: 473 and S.E: 3.1) significantly higher than students in other countries who show similar performance in mathematics, reading and science.

TIMSS results of fourth grade students from Serbia are undoubtedly more favorable than those for eighth grade students; both mathematics and science students perform significantly above the TIMSS scale average (Mathematics – Serbia: Mean: 518 and S.E: 3.5; TIMSS Scale Centerpoint: 500) (Mullis et al. 2016).

Regarding the gender differences in performance, according to PISA (2012), the average achievement of boys on a scale of mathematical literacy is 9 points more than the average achievements of girls. Although, Serbia is in the large group of countries where the education system is supporting boys in the development of mathematical competence in a greater extent, the difference falls into smaller differences. Boys are more successful than girls in the problem-solving domain (the difference is about 15 points and is statistically significant). On the contrary, on the scale of reading literacy girls performed better than boys for 46 points, which is slightly higher than the average difference recorded in the OECD countries (38 points). In terms of scientific literacy, the achievements of girls and boys are at the same level, as they were all in the previous cycle (OECD 2014a).

The socio-economic status of students explains about 8–12% of the variance depending on domains. As compared to other participating countries including the equity of the OECD countries on average, the findings of the Serbian educational system are more favourable. Serbia has above-average equity in education opportunities, and performance differences in mathematics across socio-economic groups are below the OECD average (OECD 2014b).

International Assessment Today in Serbia: Ongoing Studies and Challenges in its Implementation

Since 2001, Serbia has participated in two international assessment studies: IEA/TIMSS and OECD/PISA, with some changes in the structure and deviations in the schedule of completion (e.g. PISA 2015 cycle was skipped). For the first time, Serbia will participate in IEA/PIRLS study in the upcoming cycle. Any other international assessment study is not planned so far, not even OECD/TALIS, in which Serbia participated in one cycle, 2013.

TIMSS study is has been accomplished five times, of which the first two studies were carried out in grade 8 and the later ones in grade 4. When it comes to PISA, in the first research cycles, a basic study was conducted covering the achievements in reading, mathematics and science, while the last two studies included optional domains of problem solving (2012), financial literacy (2018) and general competence (2018).

The unstable funding was a joint challenge for both assessment studies in almost all assessment cycles. Both assessment studies rely on the Ministry of Education of the Republic of Serbia to secure funding regarding international participation fees, as well as domestic research costs. It happened more than once that the research institutions went deep into the assessment activities without finances being secured for local costs and sometimes even without signed contracts. This prevented Serbia from participating in PISA 2015.

International studies are increasingly being administered through computers. PISA 2018 is the first cycle in Serbia that was completely performed. Thus, the level of ICT equipment of schools became an issue for this study. Approximately, half of the sampled schools were not in the position to administer survey by themselves without additional computers, which were secured for this purpose by the national research center.

But the real challenge is how to make these studies and their findings regarding educational policy more influential. Hence, primary and secondary analysis identified several issues of great importance for the improvement of the teaching-learning process and further development of educational system and its quality:

1. School knowledge is inert. Students perform better on out-of-curricula domains, such as problem solving.
2. School knowledge is typically decontextualized and formal. Students perform better on formal knowledge tasks than real life tasks. (Anić and Pavlović Babić 2015)
3. There is a decline in the quality of education after the first educational cycle. (Marušić-Jablanović et al. 2017)
4. There is a huge proportion of functionally illiterate students in all domains of achievement. (Pavlović-Babić and Baucal 2013)

Low Achievement and Great Progress in Reading: Possible Explanations of Unexpected Findings

In 2009, the assessment study OECD/PISA was conducted for the third time in Serbia. The achievement data from both previous cycles (2003 and 2006) in all three examined areas (reading, mathematics and science) were disappointingly low – half of one standard deviation (or even more) below the international average (Table 18.1). In 2003, the overwhelming reaction of the public including the educational authorities and employees in the education sector was surprise, disbelief and suspicion in the findings' credibility. As early as 2006, instead of surprises, there was disappointment and skepticism regarding the quality of education, as well as the reform changes implemented in previous years. In 2009, considering that there have not been any substantial changes in the education system in the meantime, a repeat of the results from the previous cycles was expected.

Table 18.1 Mean scores and percentage of students in Serbia at each proficiency level in reading (PISA 2006, 2009, 2012)

Year	Below level 2	Level 2	Level 3	Level 4	Levels 5 and 6	Mean score
2006	51.7	28.1	16.0	3.9	0.3	401
2009	32.8	33.2	25.3	7.9	0.8	442
Change 2009 vs. 2006	−18.9	5.1	9.3	4.0	0.5	**41**
2012	33.2	30.8	23.3	10.5	2.2	446
Change 2012 vs. 2009	0.4	−2.4	−2.0	2.6	1.4	4

Source: OECD 2007, 2010, 2014

Similar to previous cycles, PISA 2009 covered about 190 schools and, respectively, 5523 students (about 95% out of planned sample). Additionally, the average achievement was statistically considerably lower than the OECD average (reading mean score: 442 and S.E: 2.4). The difference between OECD (reading mean score: 493 and S.E: 0.5) and Serbia average in reading was 51 points, which corresponds to the effect of about 1.25 years of schooling in OECD countries. As revealed in the previous cycles, the conclusion was that the education in Serbia was "slower" compared to the OECD countries. The difference is not only in the speed of achieving educational goals, but in its quality. Educational authorities and stakeholders faced the fact that, for the same number of years as their peers in OECD countries, students in Serbia gain on average competence of a lower level of complexity (Pavlović-Babić and Baucal 2013).

In summarizing the findings on the distribution of students from Serbia at different proficiency levels in reading, it can be observed that every third student after 10 years of schooling has not reached the minimum level of functional literacy (level 2). At the same time, very few students (0.8%) from Serbia managed to reach the highest levels of the reading literacy (Table 18.1).

In a nutshell, in 2009, the mean score on the reading literacy was significantly below the international average and every third student failed to reach the level of functional literacy. Still, these results represent a very significant improvement when compared with the previous PISA cycles. Specifically, in 2009, the mean achievement improved to 42 points while the percentage of students below level 2 (functionally illiterate) reduced by almost 20. This represents one of the biggest improvements between two consecutive PISA cycles.

The reduced number of students below the proficiency level 2 was not evenly distributed to all higher levels of reading, but mainly on level 3 (increase of 9.3%), followed by increase on levels 2 and 4 (Baucal and Pavlović 2010) (Fig. 18.1).

Possible Explanations of Significant Progress

The highlighted achievement data triggered two questions that should be considered, and provide the basis for certain hypotheses as answers to them. The first question would be: Why was the average achievement (and still is) significantly

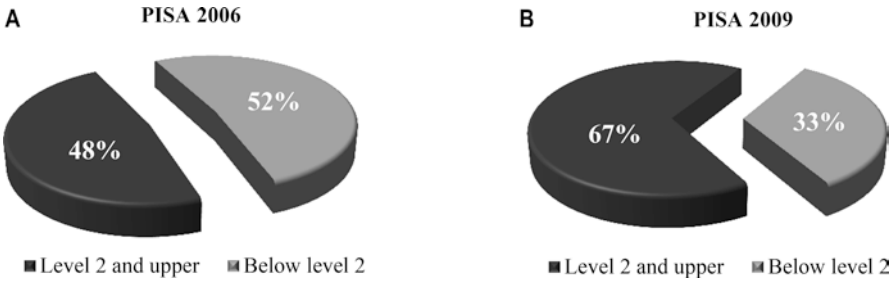

Fig. 18.1 Change in the percentage of students who have not reached proficiency level 2 in reading literacy (PISA 2006 and PISA 2009)

lower than in most OECD countries and a number of European countries? Second, which factors could explain the remarkable improvement in average achievements and the decrease in the number of students on the lower levels of achievement made in the period of the 3 years?

Why Is the Average Achievement in Serbia Below the International Average?

When formulating the answer to this question it is important to have three findings in mind stating that: (a) the average achievements of students from Serbia are similar to the achievement of the students in the neighboring countries (such as Bulgaria and Romania), (b) the differences in achievement between students in Serbia are smaller than in most other countries, and (c) a small number of students from Serbia reached the two highest proficiency levels.

What has the education in the Balkan region in common? To sum it up, the overall economic situation is weaker than in western European countries, and consequently, the investment in education is substantially lower, which is especially true when comparing the allocation of means per pupil (European Commission 2008). We believe that lower average results of students from Serbia may be partly explained by the weaker economic situation. However, earlier studies (OECD 2010; Baucal and Pavlović Babić 2010) opined that the relationship between the economic situation in the country and investment in education, on the one hand, and academic achievements, on the other hand, is not strong enough to be the only explanation.

Traditionally, the teaching–learning process in Serbia is much more knowledge oriented than competence based (European Commission 2002), while the dominant form of activities within teaching process consists of lectures in plenary with underrepresented active learning, research-based, project-based or collaborative learning (European Commission 2007; Mincu 2009; Dimou 2009).

Previous researcher findings explain the lower achievement in international assessment studies as the consequence of the typical teaching practice in Serbia which leaves small space, if any, for critical and hypothetical thinking and other higher order thinking. This explanation is supported by the fact that less than 1% of students from Serbia managed to reach levels 5 and 6; thus, the dominant form of teaching and learning in schools in Serbia does not allow even students with the highest intellectual potential to develop the highest level of competencies.

Which Factors (Could) Explain the Progress That Has Been Made?

Considering that the teaching in schools remained predominantly directed towards the adoption of academic knowledge and consequently, the knowledge expected to be demonstrated by students remained on the reproductive level; we hypothesized that this great improvement is as a result of greater engagement and motivation of the students and teachers in the PISA 2009 than previously.

This hypothesis is supported by the fact that in 2006, the Ministry of Education was largely indifferent towards PISA, which dramatically changed in 2009 when a number of positive messages and actions were addressed to schools, teachers, students and parents (e.g. a national conference on PISA assessment was held with guests from the OECD Directorate for Education; all schools received publications on this assessment study, which included accomplished items and proficiency levels' descriptions; the media regularly reported on the participating schools and students). Therefore, our assumption is that a significant number of students who had difficulty with PISA tasks in 2006, easily gave up on solving the tasks, while in 2009 these students made an extra effort to solve at least the tasks at lower proficiency levels and try to answer open-ended questions. However, the upper end of the scale motivation without adequate competencies could not improve scores (Pavlovic Babic and Baucal 2011).

Therefore, we expected that secondary analysis would confirm the following hypothesis: If the weight of items and student ability are controlled, the proportion of students who left items unanswered would be significantly higher in 2006 compared with 2009. The analysis included two independent samples of students who participated in PISA 2006 (N = 4798) and PISA 2009 assessment (N = 5523) in Serbia to determine whether the probability that students skip certain items are different in the two cycles when student ability and weight of items are controlled (Pavlović Babić 2015).

The analysis includes only those items that were used in both PISA cycles (in total, 24 items). This way, the potential effect of the differences in the items used in the two PISA cycles is controlled. Analysis was run from two PISA databases. One database consists of students' answers on open-ended reading items, the other one includes data on different student characteristics and achievements on the PISA

reading scales. It is important to note the difference between the items at the end of the brochure which can remain blank because the students were unable to reach them (missing items) and the items that the students came to, but did not answer them (skipped items). The focus was only on the latter, assuming that they reflect students' readiness to try to solve them better. The results of the analysis confirmed the expectation that the average number of skipped items in 2006 was 1.86 (8%), and in 2009 it was 1.24 items (5%). In 2009, about 75% of students managed to do all the items or to skip only 1 and in PISA 2006, this was the case in only 67%. Taking into account that the achievement of students in PISA 2006 was significantly lower (mean score: 401) than in PISA 2009 (mean score: 442), it is reasonable to assume that a higher percentage of skipped items in PISA 2006 may reflect lower student abilities. In order to validate this hypothesis, the number of skipped items in these two cycles was compared with PISA scores for students as covariates. Based on this comparison, it can be concluded that part of the differences in the number of skipped items can be explained by the achievement data, but the unexplained difference is still significant (about 0.5 items or 2% more in 2006 than in 2009). Hence, the conclusion is that the willingness of students with the same abilities to invest effort and solve the PISA item was somewhat lower in 2006 than in 2009 (Pavlović Babić 2015). This change in relationship can explain, though partially, the difference in the average achievement of students in Serbia in 2006 and 2009, suggesting that PISA students' achievement can reflect different motivational factors, such as achievement motives, attitudes towards this research and characteristics of the requirement in an item (e.g. open questions asking students for more elaborate answers), etc.

Where to Look for a Recipe for Better Achievement?

When discussing factors that can explain the students' achievement data, it is natural to expect that factors which make a difference and have the highest explanatory power act at the level of an education system. In particular, when explaining differences (progress) in students' achievements, it could be expected that the progress is as a result of system and reform measures implemented at the level of the whole education system, whether these being measures concerning the organization process of learning/teaching or its structural aspects and measures implemented during the years preceding the change registered in the achievements. As far as Serbia is concerned, the first decade of the twenty-first century was characterized by reform changes, but was predominantly short-lived. As the government was changed, the education policies and legal regulations changed also. Disappointingly, the previously described case – the progress made in Serbia in terms of reading literacy – could not be explained by factors which acted at the level of education system. Basically, significant changes in the education, either in terms of the curriculum, dominant teaching practice or in other aspects of the system, such as funding, preservice teacher education, legal regulation, the structure of the system and external

assessment practice did not occur (Pavlović Babić and Baucal 2010). Instead, factors which have been identified as responsible for improvement of achievement acted at the level of society as a whole (demographic factors, socio-economic status of family and society) or situationally (motivation of achievement).

This does not mean that the education system remains the same over time, but that the inertia of the system and the routine in the work of teachers has managed to amortize the contributions of reform initiatives, which are, again, implemented without sufficient pre-preparation or only in certain segments of the system. That is why there is less to say here about how the education system has been used in assessment data, and much more about how it can use them wisely.

What Has Been Done So Far?

Till date, the international assessment studies' conceptual solutions, ways of assessing achievements and findings have been used for some strategic goals. For example, PISA achievement data was set as an indicator in the implementation of the Poverty Reduction Strategy (the Government of the Republic of Serbia 2003). The international assessment studies directly influenced the definition of goals and the setting of priorities in the further development of education, which is visible in the Strategy of the Development of Education in Serbia, 2020 (Ministry of Education, Science and Technological Development of the Republic of Serbia 2013), as well as in law and by-law regulations (e.g. Law on the Foundations of the System of Education in Serbia, 2009 and later amendments).

Further operationalization of the educational goals defined by the Law and the Strategy was done with the development and adoption of educational standards. The process of introducing educational standards into the education system started in 2010, when standards for the end of comprehensive education (primary education, after 8 years of schooling) were adopted by the |National Educational Council. (Ministarstvo prosvete & Zavod za vrednovanje kvaliteta obrazovanja i vaspitanja 2010). This document covered 10 subjects, and was followed by the educational standards for the end of the first cycle of primary education in 2011 (4 years of schooling), for the end of the first and third cycle of primary adult education (2013), for the end of general and vocational secondary education (2016), and more recently, for the foreign languages (2017). While working on the development and implementation of these standards, the concept of standards was developed as well. It could be seen that the first standards were mostly based on content, but over time, they have become increasingly based on competences, primarily on competencies operationalized for measurement purposes in international assessment studies PISA and TIMSS. Even more noticeably, competence-based approach could be seen in the definitions of the transversal competencies for the end of primary and general secondary education. Both documents, adopted in 2013, are visibly influenced by the European Reference Framework of Key Competences for Lifelong Learning

(European Communities 2007) and international assessment studies' definition of the literacy in different domains. Unfortunately, the concept of transversal competences has an initial weakness regarding their implementation in the teaching/learning process and, consequently, on student achievements. Namely, this document strictly separates two groups of competences. First, subject competencies are developed through regular classes and are based on syllabuses. Second, transversal competences are based on the noble goal of integrating teaching content and work more deeply and actively on it, but they also have neither a syllabus nor a defined place in the teaching process. Basically, they are a sort of "educational orphans", left to the good will, sensitivity and readiness of teachers to work on them, and this is certainly not enough for a systemic impact on the achievements. Within this design of implementation, all good intentions that the concept of key competences in lifelong learning potentially carry are marginalized or lost (Papić and Garabinović 2017).

Finally, it should be understood that the final examination for all eighth-grade students (at the end of the compulsory education) is fully based on the educational standards. This examination unites two functions: it certifies the basic level of education and serves for selection when enrolling in secondary school. At the same time, the final examination is the only external examination that covers the entire generation that we have in the assessment system for now.

What Is Left: Unexploited Potentials of the International Assessment Studies

Despite the fact that achievement in national and international testing is a measure of the quality of the education system, the research findings of the international assessment studies so far have no significant (if any) influence on the education policy in Serbia. The absence of this influence is evident in the curriculum. For example, learning and teaching readership skills are planned only in the first three grades of primary school, where language teaching is significantly more focused on isolated facts than on skills and does not recognize the importance of developing readership strategies through the teaching/learning process. Therefore, one important goal of all primary findings and secondary data analysis remains, after years of implementation, informing and initiating appropriate educational policy decisions. The aspects of the education system and teaching about which international assessment studies have something to say are:

Changes in Curriculum The meaningful use of these findings would relate primarily to changes in the curriculum. However, changes of the curriculum should be essential. A comprehensive curriculum change would mean a changed attitude towards the concept of literacy based on the generic knowledge and development of higher-order thinking skills, which is set as the most general and ultimate goal of general education. This goal can be reached not only by a subject's syllabus, but also through the methods of work and general school ethos that should characterize

the climate of high expectations and a collaborative relationship between teachers and students.

Key Competencies for Lifelong Learning Reading, or reading literacy, should be given the status of interpersonal competence. This would provide a double effect; first, reading competence would get an explicit attention in the teaching process, that is not the case now and, second, this measure would explicitly promote the responsibility of every educational subject for the cultivation of reading literacy. Objective outcomes should be defined in such a way that they establish a relationship with general inter-subject competencies (the extended context of application), and the curriculum should be so formed that it clearly shows how and where each concrete subject contributes to the development of general competencies, and above all, to reading.

In-service Teacher Education An important implication of these findings relates to raising the competencies of the subject teachers to cultivate student work strategies. Pavlović (Pavlović Babić 1995) and Baucal and Pavlović (Baucal and Pavlović Babić 2010) opined that reading is not defined as an educational outcome in subject teaching, but a number of formulations of educational standards relate to reading. However, during initial education, teachers of the Serbian language do not have the opportunity to learn the text strategies or methodical approaches that would allow them to generate these strategies in working with students and this applies to a large number of foreign languages. In line with the previous recommendation, it would be logical for this recommendation to apply to all teachers, not just those who teach "language" subjects.

Teaching Methodology The next important implication concerns the method of work; more precisely, the attitude of teachers towards students' attempts to respond to more complicated questions and/or solutions to problem situations. To be precise, encouraging attempts, regardless of outcome, provides an opportunity to cultivate various problem-solving strategies and to evaluate with the guidance of teachers. These are valuable teaching situations whose potential, as it seems, is not sufficiently used in Serbia.

Finally, to answer the question from the title of this chapter, in the findings of international assessment studies, in those on achievement trends and in comparative data on other education systems as well as, in background variables that are at disposal in Serbia, there are many recipes that could inform and inspire education policy makers in the country. It is only necessary that these data are readily read and translated into practice, with sensitivity to cultural patterns and educational tradition in Serbia.

References

Anić, I., & Pavlović Babić, D. (2015). How we can support success in solving mathematical problems? *Teaching Innovations, 28*, 36–49.

Baucal, A., & Pavlović Babić, D. (2010). *Nauči me da učim, nauči me da mislim PISA 2009 u Srbiji prvi rezultati*. Beograd: Institut za psihologiju, Centar za primenjenupsihologiju.

de Landsheere, G. (1974). *IEA and UNESCO: A history of working co-operation*. Paris: UNESCO Publishing.

Dimou, A. (2009). Politics or policy: The short life and adventures of educational reform in Serbia (2001–2003). In A. Dimou (Ed.), *Transition and the politics of history education in Southeast Europe* (pp. 159–200). Göttingen: V & R unipress.

European Commission. (2002). *Key competencies*. Brussels: EU Directorate General for Education and Culture. Retrieved from: http://www.aic.lv/bolona/Bologna/contrib/EU/report_qual%20LLL.pdf

European Commission. (2007). *Science education now: A renewed pedagogy for the future of Europe*. Brussels: EUDirectorate General for Education and Culture. Retrieved from: https://ec.europa.eu/research/science-society/document_library/pdf_06/report-rocard-on-science-education_en.pdf

European Commission (2008). *New skills for new jobs anticipating and matching labour market and skills needs*. Brussels: EU Directorate General for Education and Culture. Retrieved from: ec.europa.eu/social/BlobServlet?docId=1496&langId=en

European Communities. (2007). *Key competences for lifelong learning European reference framework*. Luxembourg: Office for Official Publications of the European Communities.

Government of the Republic of Serbia. (2003). *Poverty reduction strategy in republic of Serbia.* Belgrade: Government of the Republic of Serbia. Retrieved from: https://www.srbija.gov.rs/specijal/en/20691

Havelka, N. i saradnici. (1990). *Efekti osnovnog školovanja*. Institut za psihologiju: Beograd.

Law on the Foundations of the Education System, RS Official Gazette no. 72/2009, 52/2011 and 55/2013.

Marušić Jablanović, M., Gutvajn, N., & Jakšić, I. (Eds.). (2017). *TIMSS 2015 u Srbiji*. Beograd: Institut za pedagoška istraživanja.

Mincu, M. E. (2009). Myth, rhetoric, and ideology in Eastern European education. *European Education, 41*(1), 55–78.

Ministarstvo prosvete & Zavod za vrednovanje kvaliteta obrazovanja i vaspitanja. (2010). *Obrazovni standardi za kraj obaveznog obrazovanja*. Beograd: Zavod za vrednovanje kvaliteta obrazovanja i vaspitanja. Retrived from: http://www.ceo.edu.rs/images/stories/obrazovni_standardi/kraj_obaveznog_obrazovanja/Predlog%20standarda.pdf

Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije. (2017). *Opšta, stručna i umetnička matura i završni ispit u srednjem stručnom obrazovanju*. Beograd: Ministarstvo prosvete, nauke i tehnološkog razvoja Republike Srbije. Retrieved from: http://www.dfs.rs/Pdf/VM_Koncept.pdf

Ministry of Education, Science and Technological Development of Republic of Serbia. (2018). *Progress report on the action plan for the implementation of the strategy for education development in Serbia by 2020*. Belgrade: Ministry of Education, Science and Technological Development of Republic of Serbia. Retrieved from: http://www.mpn.gov.rs/wp-content/uploads/2018/08/AP-SROS-IZVESTAJ-15jun-Eng.pdf

Ministry of Education, Science and Technological Development of the Republic of Serbia (Ed.). (2013). *Strategy for education development in Serbia 2020*. Belgrade: Čigoja Štampa.

Mullis, I., et al. (2016). *TIMSS 2015 international results in mathematics*. Boston: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Organization for Economic Cooperation and Development. (2007). *PISA 2006: Science Competencies for Tomorrow's World. Volume 1: Analysis*. Paris: OECD Publishing. Retrieved from: http://lst-iiep.iiep-unesco.org/cgi-bin/wwwi32.exe/[in=epidoc1.in]/?t2000=025279/(100)

Organization for Economic Cooperation and Development. (2010). *PISA 2009 results: What students know and can do – Student performance in reading, mathematics and science* (Vol. I). Paris: OECD Publishing. Retrieved from: https://doi.org/10.1787/9789264091450-en

Organization for Economic Cooperation and Development. (2014, February). *PISA 2012 results: What students know and can do – Student performance in mathematics, reading and science* (Vol. I, Rev. ed.). Paris: OECD publishing. Retrieved from: https://doi.org/10.1787/9789264201118-en

Organization for Economic Cooperation and Development. (2014a, February). *PISA 2012 results: What students know and can do – Student performance in mathematics, reading and science* (Vol. I, Rev. ed.). Paris, France: OECD Publishing. Retrieved from: https://doi.org/10.1787/9789264201118-en

Organization for Economic Cooperation and Development (2014b). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems* (Vol. V). PISA, OECD Publishing. https://doi.org/10.1787/9789264208070-en

Papić, M., & Garabinović. (2017, October). The position of entrepreneurship and entrepreneurial education in formal and non-formal education in Serbia. Conference: Employment, Education and eEntrepreneurship EEE17, Belgrade, Serbia. Conference Paper (PDF available). Retrived from: https://www.researchgate.net/publication/324918710_THE_POSITION_OF_ENTREPRENEURSHIP_AND_ENTREPRENEURIAL_EDUCATION_IN_FORMAL_AND_NON-FORMAL_EDUCATION_IN_SERBIA

Pavlović Babić, D. (1995). Pedagoški i socijalni uslovi kvaliteta i brzine čitanja u osnovnoj školi. *Psihološka istraživanja, 7*, 191–250.

Pavlović Babić, D. (2015). Faktori koji doprinose postignućima na PISA zadacima čitalačke pismenosti. In Radišić, J. & Buđevac, N. (Ed.), *Sekundarne analize istraživačkih nalaza u svetlu novih politika u obrazovanju* (pp. 118–131). Beograd: Ministarstvo prosvete, nauke i tehnološkog razvoja. Retrieved from: http://www.dios.edu.rs/wp-content/uploads/2015/07/sekundarne-analize.pdf

Pavlović Babić, D., & Baucal, A. (2010). Čitalačka pismenost kao mera kvaliteta obrazovanja – procena na osnovu PISA 2009 rezultata. *Psihološka istraživanja, vol., 13*(2), 241–260.

Pavlovic Babic, D., & Baucal, A. (2011). The big improvement in PISA 2009 reading achievements in Serbia. *CEPS Journal, 1*(*3*), 53–74.

Pavlović Babić, D., & Baucal, A. (2013). *Podrži me, inspiriši me PISA 2012 u Srbiji prvi rezultati*. Beograd: Institut za psihologiju, Centar za primenjenu psihologiju.

Šolić-Vojinović, M., Nastić-Stojanović, J. (2014): *Introducing tutoring Serbia. Early school leaving prevention and intervention model.* Belgrade: Tutoring Serbia. Retrieved from: http://www.wb-institute.org/meta-content/uploads/pub_Introducing-TUTORING-SERBIA-2014.pdf

United Nations Children's Fund. (2001). *Comprehensive analysis of primary education in the Federal Republic of Yugoslavia*. Beograd: UNICEF.

Vujisić-Živković, N. (2015). Constitutive discontinuity. Education and pedagogy in the socialistic Serbia (1945-1990). *Journal of Contemporary Educational Studies/Sodobna Pedagogika., 66*(2), 64–77. 27p.

Zavod za vrednovanje kvaliteta obrazovanja i vaspitanja. (2005). *Nacionalno testiranje učenika III razreda osnovne škole*. Beograd: Zavod za vrednovanje kvaliteta obrazovanja i vaspitanja.

Zavod za vrednovanje kvaliteta obrazovanja i vaspitanja. (2007). *Nacionalno testiranje učenika IV razreda osnovne škole*. Zavod za vrednovanje obrazovanja i vaspitanja: Beograd.

Zavod za vrednovanje kvaliteta obrazovanja i vaspitanja. (2017). *Izveštaj o realizaciji i rezultatima završnog ispita na kraju osnovnog obrazovanja i vaspitanja*. Zavod za vrednovanje obrazovanja i vaspitanja: Beograd.

Chapter 19
Assessment Policy and Practice of Slovenia

Klaudija Šterman Ivančič and Urška Štremfel

Following the White Paper on education (2011), one of Slovenia's most important goals in the field of education today is the establishment of a culture of quality and assessment, which is based on the concept of evidence-based policy, where participation in large-scale assessments (ILSAs) plays an important role. Beginning in 1996, Slovenia has participated in different ILSAs (PISA, TALIS, PIAAC, TIMSS, PIRLS, ICCS, and ICILS). In this chapter, we focus on PISA results, which on the one hand demonstrate that throughout the cycles beginning in 2006, Slovenian students have achieved mainly above-average results in science, mathematics and reading. On the other hand, they report rather low motivation to learn. Further, national secondary analysis results also reveal significant disparities in achievement between boys and girls, students enrolled in different educational programmes, from different socio-economic backgrounds, with different immigration backgrounds, and languages spoken at home. In this chapter, we emphasise the importance of the not-self-evident treatment of the above-average results on an international scale as the great efficacy of the national education system. At the end, the main challenges of using ILSA results to develop Slovenian educational policy and future practices are discussed.

The Education System of the Republic of Slovenia

The education system of the Republic of Slovenia is organised as a public service rendered by public and private institutions that provide officially recognised or accredited programmes. By law, public schools are secular and the school space is autonomous. The providers of public service are under supervision of the school

K. Šterman Ivančič (✉) · U. Štremfel
Educational Research Institute, Ljubljana, Slovenia
e-mail: klaudija.sterman@pei.si

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_19

inspectorate. The Slovenian education system is organised into several levels of education: Pre-school education (is optional, and encompasses the centre-based early general pre-school education and care. Children are legally entitled to a place in a kindergarten from the age of 11 months to the age of compulsory schooling); compulsory basic education (is organized in a single-structure 9-year basic school attended by pupils aged 6–15 years); upper secondary education (takes 2–5 years, typical age of students is 15–19, educational programmes include vocational, professional and general gymnasium programmes); tertiary education (includes short-cycle higher vocational education and higher education study programmes); and adult education (is marked by its considerable diversity of programmes and institutions). At the end of grades 6 and 9 of compulsory education, pupils undertake compulsory national assessment in three subjects. Tracking of students begins in upper secondary education, typically at the age of 15, after they finish grade 9. Students may choose freely among general and vocational programmes. If the number of candidates exceeds the number of places, schools may limit enrolment in the first year. In this case student's scores from grades 7 to 9 are considered, and in some cases scores on national assessment in grade 9 are also taken into account. At the end of upper secondary education, students take final exams (school leaving examination in 2- and 3-year programmed and vocational or general Matura in 4-year programmes). Matura is a national external examination which allows the students to enrol in tertiary study programmes (EACEA 2019).

Introduction to the International and National Assessment Contexts and Their History

The educational system in present-day Slovenia has a long history, which can be divided into four phases: imperialistic (until World War II), supervised (from World War II until the 1990s), sovereign (post 1991), and globalised education policy (post 2004) (e.g. Štremfel 2015). An important turning point in its development occurred in the 1990s, following Slovenia's independence in 1991. The comprehensive education reform, which occurred during that time, was characterised by a desire to break away from socialist ideological influences and get closer to the modern standards of the developed Western Europe (e.g. Gaber 2008). In addition to adopting comprehensive legislation (1991–1996) and curriculum reform (1997–1999), a major focus in Slovenia in establishing its sovereign educational system was to identify and provide the quality of education (see Table 19.1) consistent with international trends. At the same time, it served as a mirror of several issues and dilemmas about the present system and the further development of education in Slovenia (e.g. Kos Kecojević and Gaber 2011). How important it was for Slovenia to follow international trends of global standards of educational quality and achievement is also evident from the White Paper on education (1996, p. 71): "In Slovenia, one of the goals of the renewed school system is to allow achieving internationally

Table 19.1 Involvement of Slovenia in international large-scale assessments

Knowledge, skills, competence measured	Supervised policy (1945–1990)	Sovereign policy			Globalised policy (2004 onwards)
		Legislative change (1991–1996)	Curriculum reform (1997–1999)	Evaluation of the reform (2000–2004)	
Reading		RL 1991		PIRLS 2001	PIRLS 2006
Mathematics, science		TIMSS 1995	TIMSS 1999	TIMSS 2003	TIMSS 2007, 2011, 2015
		IAEP 1991			
Reading, mathematics, science					PISA 2006, 2009, 2012, 2015, 2018
Civic			CIVED 1999		ICCS 2009, 2016
Foreign language		Language Education Study (1995)			ESLC 2011
ICT			SITES M 1999	SITES M 2001	SITES 2006
					ICILS 2013
Adult skills			IALS 1998		PIAAC 2013

Notes: *CIVED* Civic Education Study, *ESLC* European Survey on Language Competences, *IAEP* International Assessment of Educational Progress, *IALS* International Adult Literacy Survey, *ICCS* International Civic and Citizenship Study, *ICILS* International Computer and Information Literacy Study, *PIAAC* Programme for the International Assessment of Adult Competences, *PIRLS* Progress in International Reading Literacy Study, *PISA* Programme for International Student Assessment, *RL* Reading Literacy Study, *SITES M1* Second Information Technology in Education Study Module 1, *SITES M2* Second Information Technology in Education Study Module 2, *TIMSS* Trends in International Mathematics and Science Study
Source: IEA (2018; https://ilsa-gateway.org/)

comparable standards at the end of the primary school". The end of comprehensive education reform (and the recognised need for evaluating its effects) in Slovenia coincided with a global paradigmatic shift towards a knowledge-based society/economy. This has in Slovenia, as elsewhere, undoubtedly concentrated the focus of education on measuring achievement and setting new standards of quality assurance, as indicated by the growing number of evaluations of educational programmes and institutions (Kos Kecojević and Gaber 2011). Many authors (e.g. Biesta 2007) believe that the shift towards outcome-centred education is closely associated with the concept of evidence-based policy-making. The empirical study (Štremfel 2013) reveals that, according to the perception of key national actors (8 policy-makers and 22 experts participating in the study), the concept of evidence-based education in Slovenia is still to be developed.

Great aspirations for following international trends in education since the beginning of its sovereignty on the one hand, and the paradigmatic shift towards outcome-centred education from 2000 onwards, on the other hand, resulted in increasing Slovenian involvement in large-scale, international assessments (ILSAs; Table 19.1).

In the 1990s, the TIMSS and PIRLS framework, which is more curriculum-based, allowed the exploration of student achievement in reading, mathematics, and

science for the periods before, during and after the reform, which is why the focus at the time was on the results of these studies. A greater incentive for Slovenia to participate in PISA in 2006 for the first time was not only Slovenia's candidature for membership in the OECD (2007–2011), but also its accession to the EU in 2004. The EU benchmark measures the percentage of 15-year-olds who fail to achieve basic levels of reading, mathematics and science literacy in PISA and encourage member states to attain the common EU goal (less than 15% of low achievers by 2020) by comparing their attainment and sharing good practices in attaining it. In the meantime, Slovenia joined other ILSAs measuring different competences and skills, such as language competence (Language Education Study, ESLC), civic competence (CIVICS and ICCS), and computer and information literacy (SITES, ICILS), as well as adult skills (IALS, PIAAC). Slovenia's involvement in various ILSAs allows international comparison of the achievements of Slovenian students and adults in different educational contexts, and also measures trends when participating in the same study in more cycles over longer periods of time. At the same time, the requirement of ILSA to achieve the technical standards of data collection, particularly at the beginning, contributed to improving the quality of Slovenian research, as a country without a strong previous tradition in this field (Štraus 2005).

How important ILSAs remain in the Slovenian educational system is also evident from the White Paper (2011, p. 24–25), which states:

> One of the important goals of Slovenian education is to ensure internationally comparable education for our pupils and students. … To achieve internationally comparable education of our students, in addition to internationally harmonized curricula and standards of knowledge, we must also achieve international harmonization of the criteria for assessing knowledge, of course with those countries that we want to compare. … At the state level, we have to clearly set and pave the way to the goal, that according to the quality of the presented knowledge, Slovenian students rank to the top that is at least the top third of the achievements of the students of the developed countries.

The next section discusses the role of ILSAs in Slovenia's current assessment framework.

International and National Assessments in Slovenia Today

Following the EU Strategic framework of Education and Training 2020, one of the most important goals for Slovenia in the field of education today is the establishment of a so-called culture of quality and assessment, based on the concept of evidence-based policy. Therefore, Slovenia is currently upgrading an existing framework of assessment and educational quality assurance, which has been built since the end of the 1990s. The framework is based on the following forms of (internal and external) systematic assessment:

- Internal self-evaluation in schools (From 2008, schools in Slovenia must carry out annual internal evaluations according to the Organization and Financing of Education Act.);
- External knowledge tests (National assessment of knowledge at the end of Grades 6 and 9, defined in the Elementary School Act (1996), General Matura and Vocational Matura, defined in the Matura Examination Act (2006)) conducted by the National Examinations Centre;
- National (2-year) evaluation studies, established by the Organisation and Financing of Education Act, defined by ministerial acts and conducted by research institutes and universities;
- External evaluation of schools defined by the School Inspection Act (1996) and conducted by the Inspectorate of the Republic of Slovenia for Education and Sport;
- External evaluation of the system by participating in different ILSAs defined as a priority in the White Paper on Education (2011) and conducted by research institutes (Educational Research Institute and Slovenian Institute for Adult Education).

Such an approach aims mainly at establishing professional cores that could support teachers in their process of empowerment in the fields of formative assessment and evaluation of their own work and their students' work and knowledge, and at assessing and improving the quality of the educational system as a whole (MIZS 2017; OECD 2016b). In the latter, participation in ILSAs plays an important role.

National results in international comparison are an important part of establishing evidence-based policy and enhancing the quality of the Slovenian educational system (White Paper on Education, 2011). Consequently, today a country participates in different ILSAs, mainly those carried out by the International Association for the Evaluation of Educational Achievement (IEA) and the Organisation for Economic Co-operation and Development (OECD), that is, OECD PISA, OECD TALIS, OECD PIAAC, IEA TIMSS, IEA PIRLS, IEA ICCS, and IEA ICILS (Table 19.1).

Currently, the main goal on a national level is to establish a systematic approach to quality assurance in education that would incorporate the advantages and common parts of all of the above-mentioned studies, as stated by Wagemaker (2014, p.13): provision of high-quality data to improve policy-makers' understanding of key school-based and non-school-based factors influencing teaching and learning, provision of high-quality data as a resource for identifying areas of concern and action, and for preparing and evaluating educational reform and development, and improvement of the capacity of educational systems to engage in national strategies for educational monitoring and improvement.

In this chapter, we discuss an example of ILSA results in Slovenia. In this context, we decided to take a closer look at PISA results and trends, because they represent an important aspect of educational policy at national and international levels.

An Example of International and National Assessment Findings: The PISA Study

Beginning already in 1995, Slovenia has participated in different ILSA studies, in the case of PISA, from cycle 2006 onwards. Throughout the cycles, Slovenian 15-year-olds mainly achieved above the OECD average in all three measured PISA domains (science, mathematics and reading; Table 19.2). The exceptions are reading achievements in 2009 and 2012, which were significantly below the OECD average. Trends in all three domains also tend to be at least stable, if not positive (OECD 2016a).

In the case of science PISA literacy, the past 9 years indicate stable, statistically significant above-OECD-average results. Although scores for Slovenia on an international science scale dropped an average of 1.5 points in every cycle, the difference did not prove to be statistically significant. The last data from the 2015 cycle also revealed that 85% (the OECD average is 79%) of Slovenian students achieve the baseline level of proficiency in science (Level 2 on the science literacy scale). It is expected that all students should attain Level 2 by the time they leave compulsory education. This way, they are able to successfully continue their secondary education and are able to tackle everyday tasks related to different literacy contexts. According to PISA 2015 data, 11% of students attained the highest level of proficiency in science, which is also above the OECD average (8%). The percentage of high-performing students has decreased by 2 percentage points from the 2006 cycle (from 12.9% in 2006 to 10.6% in 2015), but proved to be stable in the last two PISA cycles (OECD 2016a; Štraus et al. 2017).

Achievement in mathematics in Slovenia also proved to be stable and significantly above the OECD average over time. In PISA 2015, Slovenian students achieved on average 510 points on the mathematics PISA test; students in OECD countries achieved 490 points. Only European students from Estonia (520 points) and Switzerland (521 points) scored higher than Slovenian students. Between the 2012 and 2015 cycles, there was a statistically significant increase (9 points) in

Table 19.2 Average PISA performance in mathematics, science and reading in Slovenia, 2006–2015

	Mathematics		Science		Reading	
	Slovenia average	OECD average	Slovenia average	OECD average	Slovenia average	OECD average
PISA 2006	504	498	519	500	494	492
PISA 2009	501	496	512	501	483	493
PISA 2012	501	494	514	501	481	496
PISA 2015	510	490	513	493	505	493

mathematics achievement on the PISA test in Slovenia, but the average 3-year positive trend (1.7 points) did not prove to be significant. From PISA 2015 data, it is also evident that there was a decrease in the percentage of students attaining the lowest levels (below Level 2) on the mathematics literacy scale (from 20% to 16%) between cycles 2012 and 2015, while the percentage of students attaining the highest level remained stable. According to PISA 2015 data, 84% of Slovenian students achieved baseline proficiency in mathematics (Level 2; OECD average is 77%), and 13% (the OECD average is 11%) of students achieved the highest levels (Levels 5 or 6; ibid.).

Reading literacy proved to be the weakest domain of Slovenian 15-year-olds, according to PISA 2009 and 2012 results, with scores significantly below the OECD average. However, from the cycle 2009, the average 3-year trend was positive and among the highest (11 points) between participating countries. For example, in the cycle 2015, students scored significantly higher on the reading literacy scale than in 2012, that is, 24 points (481 points in 2012 and 505 in 2015), and also above the OECD average (493 points). According to PISA 2015 data, 85% (vs. 80% for the OECD) of Slovenian students achieved the baseline proficiency level (Level 2) in reading, which is 6 percentage points more than in 2012, and 9% (vs. 8% for the OECD) of students achieve the highest levels (Levels 5 and 6) on the reading literacy scale. As such, Slovenia is the only one of the participating countries where the percentage of high-proficiency students in reading increased between 2012 and 2015, and the percentage of low-proficiency students decreased (ibid.).

On the other hand, national data demonstrate that there are significant differences in the achievement according to achievement predictors, for example, gender, educational programme, socio-economic status, immigrant background, language spoken at home and motivation to learn.

PISA 2015 results indicate that Slovenia is still among the countries with the largest significant gender gap in reading achievement (43 points) in favour of girls (528 points vs. 484 points). Results also reveal significant differences in science achievement between girls and boys, also in favour of girls (516 points vs. 510 points). In mathematics, boys on average performed slightly better than girls (512 points vs. 508 points), but the difference didn't prove to be statistically significant (Štraus et al. 2017).

Further analysis of national data (e.g. Šterman Ivančič and Puklek Levpušček 2018) also revealed significant differences in achievement in all three domains according to the student's educational programme. Results of the secondary analysis demonstrate that, for example, students from general gymnasium programmes scored on average 584 points on the science literacy scale, students from technical educational programmes scored 499 points, and students from vocational-educational programmes scored 418 points. The difference in scores between students in general gymnasium and vocational programmes is approximately 160 points, which corresponds to approximately 5 years of schooling.

PISA 2015 national results also revealed that students from different educational programmes vary according to their reported socio-economic background: for example, the index of socio-economic status for students in general gymnasium programmes has a value of 0.53 and the value of the index for students from

vocational-educational programmes is −0.61. Furthermore, PISA 2015 results indicate that students from the bottom quarter of index of socio-economic status on average scored 471 points on PISA science scale, and students from the top quarter on average scored 560 points. The gap between the two quarters corresponds to 88 points, which is equal to the gap identified in the OECD average. Such results could therefore indicate that the differences in science scores between the educational programmes in Slovenia could at least partially result from the differences in students' socio-economic background, with the difference between the achievements of students with different socio-economic background not significantly different from the average difference in other OECD countries.

According to PISA 2015 data, there are also significant differences in science achievement between students with a migrant background (7.8% of participating students) and non-migrant students (92.2%). On average, students with a migrant background achieved 449 points on science PISA 2015 test, which is 71 points lower than their peers with non-migrant background (520 points). The difference in the proportion of low-performing science students with a migrant background and non-migrant students is 7.8% in favour of non-migrant students, which is also relatively high according to the EU average (European Commission 2017).

Significant differences in science achievement were also found between students whose language at home is Slovenian and students for whom it is not. Students who speak Slovenian at home on average scored 88 score-points higher than students whose mother tongue is not Slovenian, which corresponds to almost 3 years of schooling (European Commission 2017).

Already from the PISA 2009 cycle on, the results demonstrate that, despite the fact that Slovenian 15-year-olds achieve high scores in all three measured domains, they report rather low learning motivation: for example, the value of the index of enjoyment in learning science for Slovenia was significantly below the average (−0.36); similar results were reported for the index of interest in broader science topics (−0.32); students in PISA 2015 also reported below-OECD-average instrumental motivation to learn (−0.45). The same was found for the motivation to read (In PISA 2009, students reported on average below-average enjoyment in reading (−0.20).) and motivation to learn mathematics in the PISA 2012 cycle (−0.03; e.g. Kozina and Štraus 2017; Šterman Ivančič 2017; Šterman Ivančič and Puklek Levpušček 2016). TIMSS 2015 also yielded similar results in 2015, especially regarding low motivation to learn science (Japelj Pavešić 2017). The contradiction between high achievement results, on the one hand, and low motivation to learn, on the other hand, has been widely discussed by national educational experts and the research community, and is still one of the important topics when addressing ILSA results in Slovenia.[1]

[1] At this point it is also important to note that these results are derived from the comparison of national non-cognitive results to the OECD average. Further analysis in comparing different subgroups according to non-cognitive indicators with students' achievement within a country are therefore of crucial importance in the future.

We can conclude that, despite the above-average results of Slovenian 15-year-old students in science, reading, and mathematics, further analysis of PISA data is crucial to actually understand the nature of the results and, most importantly, that the above-average achievement on an international proficiency scale should not be self-evidently treated as evidence of the great efficacy of the national educational system. In the next chapter, we discuss further the role of ILSA results in the improvement in students' academic achievement and the challenges of using ILSA results to develop Slovenian education policies and practices.

How to Go Further? A Critical Discussion of Assessment Policies, Practices and Results

ILSA results reveal that disparities in student achievement according to gender, educational programme, socio-economic background, immigrant status, language spoken at home, and learning motivation present an important challenge to the Slovenian educational system and respecting equity as one of the main principles of the White Paper (2011).

Based on the case studies (e.g. Japelj Pavešić 2013; Klemenčič 2010; Štremfel 2013; Šimenc 2012), these and other ILSA results play an important role in the assessment and improvement of academic achievement in Slovenia. The results of case studies have, from different research perspectives and taking the international framework of particular ILSA into account, identified the following influences of ILSAs on the process and content of education policy in Slovenia.[2] The findings (Japelj Pavešić 2013) suggest that TIMSS results represented an argumentation for some directly and indirectly curricula and syllabus changes over the years.[3] These were accompanied by intensified teacher trainings and changes in teaching and learning process (e.g. introduction of experimental learning and use of TIMSS type of exercises in the teaching process) (ibid.). Šimenc (2012) also found a relatively high content match among civic and citizenship curricula (1999, 2011) and ICCS 2009 framework. However, according to the authors, we cannot claim that CIVED 1999 and ICCS 2009 directly affected these curricula.

Klemenčič (2010) reported that PIRLS 2001 results presented one of the foundations for forming the National Literacy Strategy. Štremfel (2013) identified the influences, which the below-average results of Slovenian students in PISA 2009

[2] It should be pointed out that national policy-making (including curricula changes) is a result of different intertwined factors and cannot be solely and directly attributed to the ILSA influences (e.g. Klemenčič 2010; Štremfel 2013).

[3] For example, TIMSS 2003 provides some kind of external evaluation of the reform of the educational system happening in Slovenia in years after its independence (1991). It enabled comparisons of mathematics and science achievements of students enrolled in the old 8-year and new 9-year elementary schooling. The revealed weaknesses of the new 9-year curriculum resulted in its immediate changes.

caused in the education policy process. These involve the organisation of national conferences, regional discussions and workshops, targeted projects for improving the level of reading literacy among Slovenian students, and intensive media attention to below-average results. Adam (2014), Štefanc (2008) and Vezjak (2014) exposed that involvement of Slovenia in international integrations in the field of education (including, but not solely, in ILSAs), resulted in some unintended changes in the national education policy (e.g. orientation to the economic dimension of education).

We can assume that although Slovenia as a new (post-socialist) EU member state has been very receptive to comparisons with the developed West (ideational pressure), it has also been confronted with institutional difficulties in translating these ideas into a national context, because of institutional and organisational constraints (e.g. lack of sources and researchers dealing with in-depth secondary analysis of the ILSA results) (Štremfel 2013). The main challenges to using ILSA results to develop a Slovenian education policy and practices in the future are:

– *Systematic planning and involvement in the forthcoming ILSA*, which does not depend solely on the (non)availability of public funding for involvement, but on expert and political consensus (e.g. which competences of Slovenian youth are strategically important for Slovenia to compare internationally, which trends in knowledge of Slovenian students is it important to follow, etc.). The lack of public funding has already initiated the debates of (non)involvement of Slovenia in different ILSAs (e.g. ICILS in 2018).
– *Responding to the ILSA results not only when faced with below-average results.* The important power the ILSAs have over participating states lies in the fear of being below average, which is becoming more and more common in these times of increasingly fierce competitiveness in the globalised world (e.g. Ozga 2003). Silova (2012) explains this fear is even more evident in post-socialist states, where any deviation from the Western "norm" is immediately reflected in the emerging narratives of "crisis", "danger", and "decline". The research evidence (e.g. Štremfel 2013) confirms the increasing (political) attention to ILSA results for below-average results in Slovenia. This, among others things, resulted in paying more attention to and investing in additional secondary analyses of PISA results to better understand the broader social, psychological, and economic contexts of the results. Kodelja (2005) points out that excessive emphasis on poor results as a form of political response to unsatisfactory results is not necessarily a bad thing, providing they are justified. He believes that they may even significantly contribute to improving the situation, as they shape a social climate that favours change. However, when they are based on simplifications, sweeping generalisations, and hasty conclusions, and there has been no lack of these in Slovenia in the past, they certainly do not work in favour of either identifying the real causes of the current situation or the search for solutions to improve it.
– *Not using the ILSA results for politically motivated changes.* The existence of ILSA (neutral) expert data does not ensure that these data are not used for politically motivated changes. Empirical evidence (Klemenčič 2010; Kodelja 2005)

actually reports on the use of ILSA data for politically motivated changes. In addition, the empirical study (Štremfel 2013) reveals that 63% of policy-makers, 81% of experts, and 84% of principals participating in the study agreed with the statement "International comparative assessment studies in Slovenia are often used as an argument for politically motivated changes in the field of education".

– *Developing advanced research infrastructure for in-depth secondary analysis of the ILSA results.* Slovenian representatives emphasise that the relevant skills are required not only for the purposes of conducting ILSA, but for the interpretation and contextualisation of student results. For an increasingly large amount of data, made possible by modern technology in conjunction with a not fully developed culture of this type of educational research and scarcity of human resources, the in-depth interpretation of these data remains a challenge for Slovenia and several other participating states.
– *Developing evidence-based education.* In accordance with the theoretical assumption of the governance of problems, increasing participation in ILSAs allows a wider identification of the weaknesses and shortcomings of a national educational system. The case study (Štremfel 2013) confirmed that Slovenian actors (policy-makers and experts) believe that the results of ILSAs allow identification of national policy problems when it comes to Slovenia's below-average results. However, inadequately perfected institutional structures for processing and interpreting data are insufficient pathways for developing country-specific solutions to perceived policy problems.

In any case, an evidence-based policy approach that also acknowledges international assessment results when tackling different issues in the educational arena evolved over time in Slovenia and is still evolving in an encouraging way. Still, much effort will be required to achieve a status where ILSA results will be acknowledged and incorporated in system-level decision-making as a whole, not only about the final achievement in mathematics, reading or science scale. Areas of improvement in this manner are being further discussed.

It seems that an important step towards addressing the above-identified challenges presents the new framework for identifying and ensuring quality in the field of education. The new framework aims to unite different existing (internal and external) approaches of monitoring and evaluating educational institutions and system in single comprehensive model. Among other things, the new framework established the so-called Coordination and Analytical Centre of quality assurance at the Ministry of Education, Science and Sport (the Office for the Development of Education), which is responsible for preparing a joint evaluation of the educational system (partly at the annual level and summary quality report, presumably for 3 years). It is foreseen that it will significantly contribute to a more systematic, higher quality, and more comprehensive planning of measures and development policies in the field of education on the basis of expert data and evidence (including from ILSAs; MIZS 2017).

References

Adam, A. (2014). Analiza predloga sprememb zakona o gimnazijah s perspektive odnosa med šolo in ekonomijo (Analysis of the amendments to the law on gymnasiums from the perspective of the relationship between school and economy). *Vzgoja in izobraževanje, XLV*(3), 15–21.

Biesta, G. (2007). Why »what works« won't work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory, 57*(1), 1–22.

EACEA. (2019). *Slovenia overview*. European Commission: EACEA National Policies Platform. Retreived from: https://eacea.ec.europa.eu/national-policies/eurydice/content/slovenia_en

European Commission. (2017). *Education and training monitor 2017: Slovenia*. Luxembourg: Publications Office of the European Union. Retrieved from: https://ec.europa.eu/education/sites/education/files/monitor2017-si_en.pdf

Gaber, S. (2008). Snapshots of policymaking in a changing environment. In B. Chakroun & P. Sahlberg (Eds.), *ETF yearbook 2008: Policy learning in action* (pp. 101–106). Torino: European Training Foundation.

Japelj Pavešić, B. (2013). TIMSS in Slovenia: Reasons for participation on 15 years of experience. In L. S. Grønmo & T. Onstad (Eds.), *The significance of TIMSS and TIMSS advanced, mathematics education in Norway, Slovenia and Sweden* (pp. 51–90). Blindern: Akademika Publishing.

Japelj Pavešić, B. (2017). Who likes learning mathematics and science in school? *School Field: Journal for theory and research in the field of education, 28*(5/6), 55–86.

Klemenčič, E. (2010). The impact of international achievement studies on national education policymaking: The case of Slovenia – How many watches do we need? In A. W. Wiseman (Ed.), *The impact of international achievement studies on national education policymaking* (pp. 239–266). Bingley: Emerald.

Kodelja, Z. (2005). Komparativne edukacijske raziskave in šolska politika (Comparative education research and school policy). *Šolsko polje, 16*(3–4), 211–226.

Kos Kecojević, Ž., & Gaber, S. (2011). *Kakovost v šolstvu v Sloveniji* (Quality in education in Slovenia). Retrieved from http://www.solazaravnatelje.si/isbn/978-961-6637-32-9.pdf

Kozina, A., & Štraus, M. (2017). Relationship between academic achievement as measured in the PISA study and wellbeing indicators: Preliminary findings. *School Field: Journal for Theory and Research in the Field of Education, 28*(5/6), 185–212.

MIZS. (2017). *Nacionalni okvir za ugotavljanje in zagotavljanje kakovosti na področju vzgoje in izobraževanja* (National framework for identifying and ensuring quality in the field of education). Retrieved from http://www.mizs.gov.si/si/delovna_podrocja/urad_za_razvoj_in_kakovost_izobrazevanja/sektor_za_razvoj_izobrazevanja/ugotavljanje_in_zagotavljanje_kakovosti_v_vzgoji_in_izobrazevanju/

Organisation for Economic Cooperation and Development. (2016a). *Volume I results: Excellence and equity in education*. Paris: OECD Publishing. Retrieved from: https://read.oecd-ilibrary.org/education/pisa-2015-results-volume-i_9789264266490-en#page1

Organisation for Economic Cooperation and Development. (2016b). *Education policy outlook: Slovenia*. Paris: OECD Publishing. Retreived from: http://www.oecd.org/slovenia/Education-Policy-Outlook-Country-Profile-Slovenia.pdf

Ozga, J. (2003). *Measuring and managing performance in education*. Edinburgh: University of Edinburgh. Centre for Educational Sociology (CES) Briefing No. 27.

Silova, I. (2012). Contested meanings of educational borrowing. In G. Steiner-Khamsi & F. Waldow (Eds.), *Policy borrowing and lending in education* (pp. 229–245). London: Routledge.

Šimenc, M. (Ed.). (2012). *Razvoj državljanske vzgoje v Republiki Sloveniji* (Development of civic and citizenship education in Slovenia). Ljubljana: Pedagoški inštitut, Digitalna knjižnica, Dissertationes 22. Retrieved from https://www.pei.si/ISBN/978-961-270-146-8.pdf

Štefanc, D. (2008). Ideje neoliberalizma v procesih rekonceptualizacije obveznega splošnega izobraževanja: nekatere teoretske poteze in praktične implikacije (Neoliberal ideas in the

processes of reconceptualization of compulsory general education: some theoretical moves and practical implications). *Sodobna Pedagogika, 3*, 10–31.

Šterman Ivančič, K. (2017). Šolska klima in medvrstniško nasilje v srednjih šolah: raziskava PISA 2015 (School Climate and Bullying in High-Schools: PISA 2015 Study). *School Field: Journal for theory and research in the field of education, 28*(5/6), 157–183.

Šterman Ivančič, K., & Puklek Levpušček, M. (2016). Motivational goals and academic performance from the perspective of students' perceived quality of relationship with their class teachers at the start of the upper secondary education level. *School Field: Journal for Theory and Research in the Field of Education, 27*(1/2), 113–137.

Šterman Ivančič, K., & Puklek Levpušček, M. (2018). *Individual and teacher-level predictors of student achievement: PISA 2015*. Presentation from EARA 2018. Ghent: European Association for Research on Adolescence. Retrieved from https://eara2018.eu/pdf/Day%202.pdf, https://www.eara2018.eu/search.php

Štraus, M. (2005). International comparisons of student achievement as indicators for educational policy in Slovenia: Evaluating students' achievements. *Prospects, 35*(2), 187–198.

Štraus, M., Šterman Ivančič, K., & Štigl, S. (Eds.). (2017). PISA 2015: Program mednarodne primerjave dosežkov učencev in učenk: nacionalno poročilo s trendi dosežkov med leti 2006, 2012 in 2015 ter primeri naravoslovnih nalog. (PISA 2015: National Report With Trends Between 2006, 2012 and 2015, and Science Units Examples). Ljubljana: Educational Research Institute.

Štremfel, U. (2013). *Nova oblika vladavine v Evropski uniji na področju izobraževalnih politik* (New modes of European Union governance in the field of education policies). Doctoral dissertation. Ljubljana: Faculty of Social Sciences, University of Ljubljana.

Štremfel, U. (2015). Slovenian education policy in the EU context. In D. Lajh & Z. Petak (Eds.), *EU public policies seen from a national perspective: Slovenia and Croatia in the European Union* (pp. 267–278). Ljubljana: Faculty of Social Sciences.

Vezjak, B. (2014). Neoliberalizem v šolstvu. *Uvodnik, Časopis za kritiko znanosti, XLII*(256), 7–12.

Wagemaker, H. (2014). International large-scale assessments: From research to policy. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Statistics in the social and behavioural sciences series (Handbook of international large-scale assessment. Background, technical issues, and methods of data analysis)* (pp. 11–36). Boca Raton: CRC Press.

Chapter 20
Monitoring of Student Achievement in Spain

Alejandra Tiana

The origin of the monitoring of student achievement in Spain dates back to the 1980s, even if it was organised as a systematic activity a decade later. Since then, the system has essentially retained its institutional structure and modus operandi, despite changes in approach and in the nature of the tests applied. The system has two components: a set of national studies involving Spain's 17 regions (labelled as autonomous communities in Spanish) and a series of international studies in which Spain has played an active role since the 1990s. Whereas the approach and characteristics of the former have changed over the years—making trend studies more difficult—the continuity and comparative nature of the latter have made them a crucial element for diagnosing the performance of the Spanish education system, despite discrepancies over data interpretation. This chapter begins by outlining developments in the monitoring of student achievement in Spain and its current status, before going on to analyse the challenges posed and potential directions for future development.

Introduction to the Spanish Education System

The core of the Spanish education system is 10-year basic education. It is compulsory and free of charge for all children from 6 to 16. It covers primary education (six grades) and compulsory secondary education (Educación Secundaria Obligatoria—ESO, four grades). The structure of these two levels is the same for all the country, and there is a national core curriculum and a defined set of competencies, even if autonomous regions are responsible for the organisation and operation of schools. The system is a comprehensive one from grade 1 to 10, without streaming students

A. Tiana (✉)
Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain
e-mail: atiana@edu.uned.es

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_20

and leading to a single diploma of lower secondary education. Grade 10 has a counselling purpose, introducing some optional subjects with different orientations. Comprehensiveness is balanced with a personalized approach, paying attention to personal differences among students, introducing curricular and organisational measures to address different educational needs and offering optional subjects or ways adapted to the personal projects and expectations. Private schools offering these ten grades can receive public funds if they meet some legal requirements concerning admission's policy, parents' participation and no compulsory fees. In total, about 33% of primary and lower secondary students attend private schools (less than 5% attend non-subsidised ones).

Spanish law puts special emphasis on the educational purposes of childhood attention and care. This level is composed of two cycles: 0–3 and 3–6. And after basic education, students getting the lower secondary diploma have two different ways to continue their studies: upper secondary education (Bachillerato, two grades) or vocational training. Its main aim is to offer really equal opportunities, giving every citizen the possibility to continue their studies regardless of their socioeconomic situation. This implies the existence of a solid national grants system. It also implies stressing the aim of making lifelong learning real in a national context not always sensitive to it.

A Brief History

The history of the monitoring of student achievement in Spain dates back to the early 1980s, when an ambitious reform of primary and secondary education was set in motion, generating a lively debate about its impact on student achievement. In order to be able to respond to the debate surrounding the effects of this reform, the Ministry of Education and Science developed an evaluation programme, which produced three public reports (Centro de Investigación y Documentación Educativa 1988, 1990, 1992; García Sánchez 2011), and undertook diverse secondary analysis on the data obtained. In parallel, it was decided that in 1988 and 1990, Spain would take part in the International Assessment of Educational Progress (IAEP), promoted by the US National Center for Educational Statistics (USA), and in 1991 the Reading Literacy Study, promoted by the International Association for the Evaluation of Educational Achievement (IEA), with a view to acquiring comparative international data on performance in key subject areas. Completion of these national and international projects also required the creation and training of a technical team in charge of implementing new projects to assess student achievement.

Building on these pioneering initiatives, 1990 saw the creation of the *Instituto Nacional de Calidad y Evaluación* (INCE), which began operating in 1993. The institute was entrusted with designing and implementing a general model to monitor the Spanish education system, including assessing student achievement in the different subjects on the curriculum, and gauging the effects of the reforms undertaken. The new institute's role also included coordinating Spain's participation in

the projects promoted by international organisations like the IEA, OECD and UNESCO, in whose work it became actively involved. This marked the beginning of education system evaluation in Spain.

The education system's evaluation model implemented in Spain in the 1990s can only be properly understood in the context of the political decentralisation process that took place following the end of the Francoist regime and the adoption of the Spanish Constitution in 1978. Spain's political organisation into 17 different regions, which receive the name of autonomous communities,[1] each with a high degree of political and administrative autonomy and its own parliament and regional government and budget, has been reflected in an education system that has a common structure, a national core curriculum and state qualifications, but which is highly decentralised in terms of organisation and daily operation. Although Spain does not define itself as a federal state, its level of educational decentralisation is similar to that of many countries with this political structure, and autonomous communities are similar to the units of a federal state (Länder, provinces or states). Naturally, the evaluation of a system of this kind—with limited national responsibilities in this field—in turn requires highly decentralised structures based on cooperation between regions. This gave rise to the above-mentioned INCE, designed as a state body but within a framework of close cooperation with the autonomous communities. The regions were involved in the INCE's organisation and operations through its governing board and under the supervision of the conference of regional education ministers. They also created their own institutes, agencies and assessment bodies, forming a state network.

Since the 1990s, the INCE's institutional structure and modus operandi have remained essentially the same, despite modifications in approach and in the nature of the tests applied (Tiana 2014). The institute has also changed its name and is currently known as the *Instituto Nacional de Evaluación Educativa* (INEE).[2] Despite the new title, its operating style, cooperative character, management structure and working programme have remained broadly similar.

Over the last 25 years, its main line of work has involved assessing student achievement in primary education and compulsory secondary education (ESO in Spanish and equivalent to lower secondary), usually focused on the subjects of mathematics, Spanish language, sciences and social sciences. In some cases, there has also been an assessment of English language, ICT skills and certain other areas, although to a lesser extent than the above. A second line of work has sought to ensure Spain's participation in diverse international assessment projects. These include those promoted by the IEA, and in particular the Third International Mathematics and Science Study (TIMSS, later on Trends in International Mathematics and Science Study) and the Progress in Reading Literacy Study (PIRLS), and those sponsored by the OECD, including the Programme for

[1] Both names (region and autonomous community) are used as synonyms in this text.

[2] The INEE description, studies and publications can be consulted at www.educacionyfp.gob.es/inee.

International Student Assessment (PISA), the Programme for the International Assessment of Adult Competencies (PIAAC) and the Teaching and Learning International Survey (TALIS). It is worth noting that Spain has been very active at the international level, having taken part in every cycle of PISA (2000, 2003, 2006, 2009, 2012, 2015, 2018), five cycles of TIMSS (1995, 2003, 2011, 2015 and 2019) and four cycles of PIRLS (2001, 2006, 2011, 2016). A third line of work has sought to draw up a state system of indicators, linked to those published by the OECD in *Education at a Glance*, although with certain national specificities. This indicator system is regularly updated, and a complete report is published every 2 years. These three lines of action comprise the INEE road map, following the guidelines set out by its predecessors since the 1990s.

Despite the existence of such basic continuity, the years of the current twenty-first century have seen certain changes in the preparation of national tests and their moments of application. With regard to the nature of the tests themselves, until 2006 these were focused on the core curriculum defined at state level and essentially geared to studying the knowledge acquired and skills attained in certain key areas at the end of primary education and ESO. As of 2006, the emphasis placed on key competences as the coordinating element of the curriculum (Tiana et al. 2011) was reflected in the development of a new test model based on the assessment of these skills. It was of diagnostic design (Roca 2013), as a result of which it was applied 2 years before the end of each stage (fourth year of primary education and second year of ESO), with a view to introducing measures to improve and consolidate the formative aspect of the assessment. Finally, as of 2013, tests were reintroduced at the end of the different stages to serve as an individualised assessment of student achievement, with implications for their school records, although this decision has generated considerable controversy, as we will see below.

The Current Situation

Today, student achievement in Spain is monitored in two different ways. The first involves the international tests referred to above. PISA 2018 has recently been applied, and its findings were published in December 2019. There are plans for Spain to continue to participate in subsequent editions of the project. A specific aspect of this participation is that the state sample corresponding to the whole of Spain is enlarged and stratified at regional level in order to obtain representative and balanced data from each of the 17 autonomous communities. That decision, keeping nevertheless the Spanish sample in order to allow cross-national comparisons, enables cross-regional comparisons and with other participating countries (it is necessary to underline that national tests are designed in such a way that they do not allow cross-regional comparisons). Other international studies sponsored by the OECD in which Spain takes part include TALIS and PIAAC. Of the studies promoted by the IEA, Spain continues to take part in PIRLS (the last cycle was in 2016) and in TIMSS (the latest round is 2019), although the latter only covers the fourth

year of primary education. These international evaluations are considered a very valuable instrument to monitor Spanish student achievement from a comparative international perspective, as they contribute with a cross-regional information not provided by national tests.

The second type of tests is national or regional, the nature of and approach to which have, as we have seen, changed over the years. Tests are currently applied to students in the third and sixth year of primary education and to those in the fourth year of ESO, as set out in the LOMCE Education Act (*Ley Orgánica para la Mejora de la Calidad Educativa*—Organic Law for the Improvement of Education Quality) adopted in 2013. The primary education tests are intended, according to the law, as monitoring and guidance, and serve to provide families and schools with information, whereas those at the end of ESO are thought as high stake and have academic consequences, as they determine the acquisition or otherwise of the final qualification of basic education. In fact, they were not conceived as monitoring instruments but as tools for awarding diplomas. However, in view of the controversy generated by the adoption of the LOMCE Act, which was fiercely opposed by many political organisations and professional associations, the government decided to postpone the entry into force of the individualised tests at the end of the primary and lower secondary stages and their consequences for receiving an academic qualification. Thus, at the current time, all the tests that monitor student achievement, at both educational stages, are diagnostic in nature, without academic repercussions, and are applied by the regions to samples of students on an annual basis. The autonomous communities are responsible for preparing these tests, in keeping with general guidelines, in the case of primary education (but not for lower secondary—ESO), agreed with the education ministry. The communities are also responsible for their application and correction, for which they use their own assessment bodies and education staff (such as teachers from other schools or supervisors) or specialist companies. They are also in charge of compiling results and notifying schools. With that design, they do not provide cross-regional comparative information and are mainly addressed to promote schools' improvement plans. So, they are not the same as usual national tests applied in other countries.

Main Findings, Inputs and Debates

As we can see, Spain has developed quite a comprehensive system for monitoring student achievement in the last 30 years. The various national, regional and international tests that have been regularly conducted provide thorough and diversified information on what students learn. Overall, it can be said that this system depicts quite an accurate picture of the status of Spanish education. However, in view of the nature of the tests applied, their time course, approach and territorial cover, they are not all equally useful, nor is their media, political or educational impact the same.

If we begin by analysing national tests, it should be underlined that, since the 1990s, tests of student achievement have been prepared, applied and studied on a

regular basis, in 3- to 4-year cycles. But these tests have changed over time and have sometimes taken a different course in different autonomous communities. This variability has made it difficult to ascertain achievement trends for the country as a whole in a systematic way. Instead, the reality has been a series of applications, with obvious discontinuity and divergences, which has enabled us to monitor the education system at specific moments in time but has not provided us with a rigorous record of long-term progress.

The 1990s saw the start of tests adapted to the curricular model set out by the LOGSE Education Act (*Ley Orgánica de Ordenación del Sistema Educativo*—Organic Law on the General Organisation of the Education System) of 1990. This model established a common core curriculum to the whole of Spain which accounted for at least 55% of the curricular content of autonomous communities and schools. Moreover, the skills that students should acquire at different moments in their schooling were defined in a constructivist framework. The (sampling) tests conducted for primary education in 1995, 1999, 2003 and 2007 and for ESO in 1997 and 2000 were adapted to this curricular model and made it possible to monitor trends in Spanish student achievement over these years, more solidly grounded in the case of primary education, which were applied on four occasions, and less so for ESO, which were only applied twice. The relevant national reports[3] attracted attention at the time of publication, but the difficulty of establishing conclusive trend analysis reduced their appeal. Furthermore, the underrepresentation in these studies of the ESO, the stage that generated most debate, contributed to the public's fleeting interest in the reports and their limited impact in the media.

After the LOE Education Act (*Ley Orgánica de Educación*—Organic Education Law) was passed in 2006, the focus of the tests changed, as they became diagnostic in nature and began to measure student acquisition of key competences. The common national curricular model remained in force but introducing the reference to key competences to be acquired by students during their basic education, in line with a trend that was spreading in Europe at that time (Halász and Michel 2011). This new model, however, despite including a long-term process with cyclical tests (Instituto de Evaluación 2009), was only applied in primary education in 2009 and in ESO in 2010. As a result, the new model lacked both continuity with the previous one and the subsequent consolidation required to obtain data on trends in competence acquisition. As a result, the public and educational impact of these tests was again limited.

The adoption of the LOMCE Education Act in 2013 once again changed the focus of assessment tests. On the one hand, as we have seen, these were devised as individualised student tests, with academic consequences in some cases, although the entry into force of this latter aspect has been postponed indefinitely. On the other hand, the act changed the curricular model, modifying the previous conception of common subject areas and adding achievement standards for all educational stages

[3] They are available at www.educacionyfp.gob.es/inee/evaluaciones-nacionales/publicaciones-anteriores.html.

and subjects, which were to serve as the basis for the preparation of the new tests. Additionally, the tests were to be conducted yearly, exerting considerable pressure on assessment bodies. Finally, responsibility for the tests has been placed mainly in the hands of the autonomous communities, making the compilation of statewide data more complicated, both with regard to the specific point in time and in terms of trends and progress.

Nevertheless, this latter phase has had an interesting effect, albeit one restricted to certain autonomous communities. Since they were initially designed as census-based tests, some regions are applying them yearly to all the schools in their region, which enables the production of targeted reports for each school, a practice that has spread at this recent stage. Although the data received by schools is limited, it does enable them to ascertain trends in their performance indicators and to adopt plans for improvement. Thus, for instance, the education authorities in Catalonia draft an annual report for each school which provides indicators including average outcomes obtained by their students in each of the areas assessed, the mean data for the schools in their district, town and in the whole of Catalonia. This means any school can discover how its situation compares to similar schools, how its outcomes have evolved and what measures it should adopt to improve them. Similar initiatives have been undertaken in Madrid and Andalusia. In this way, the monitoring of student achievement has become a key factor to assess schools in a different way from previous stages (Tiana 2018), albeit the explanatory capacity of the tests, identification of trends and national comparison is limited, which in turn has reduced their media and political appeal.

Given this situation, it has been the international studies that have attracted most attention, both in the media and in political and academic circles. Their advantage over purely national studies lies in two relevant traits: firstly, the potential for comparison they introduce, particularly after incorporating all the autonomous communities in a targeted manner, with their own representative and stratified samples, and secondly the regularity of their application, which has been sustained over time (in some cases, more than 20 years). As a result, PISA, like PIRLS and TIMSS, although particularly the former, permits the monitoring of changes in performance trends and comparison between regions and other countries. It is necessary to underline that given the number and characteristics of countries participating in those studies, PISA has been much more influential in Spanish debates about education than IEA studies.

Nevertheless, these studies, and once again in particular PISA, have had a threefold impact that has gone beyond analysis of the situation in the education system (Tiana 2017). Firstly, PISA has had a major impact in the media, being Spain one of the countries to have shown most interest in sharing and discussing its findings (Ferrer Julià et al. 2006; Fernández-Aliseda Garrido 2016). Secondly, it has had considerable impact on the political discourse surrounding education (Choi and Jerrim 2016), as its analyses have been used to justify some educational reforms, even with important political debates. And thirdly, it has also had repercussions in the area of academia and research, as it has been used as a basis for professional discussions about curriculum and teaching and learning practices (Pereyra et al.

2011; Carabaña 2015). The impact of the first two has been particularly striking, and both have been harshly criticised for their frequently distorting effect. The specialists quoted above consider the media to have interpreted many times PISA's findings in an incomplete and biased manner, taking them out of context and devaluing the work done by highlighting almost exclusively more negative aspects and failing to check the data with experts. Meanwhile, in the political arena, PISA has been used by diverse stakeholders to support their particular thesis on the situation and problems in the education system and to defend their respective reform projects. Behind this phenomenon is undoubtedly the strong political charge PISA entails and the authority universally granted to the OECD.

With regard to the debate generated by PISA in Spain, it is worth noting that some of the specialists who have worked most on this data conclude that the findings for Spain are roughly comparable to those of other OECD countries (although slightly below average in absolute terms, Spain comes closer if the socioeconomic and cultural index of the students tested is taken into account) and that this is achieved with considerably fewer resources than other countries (thus indicating remarkable efficiency). The data also show one of the lowest rates of inequality by sex and social class (in a notable indication of equality). On the other hand, there are significant differences between Spanish students and immigrants, and there is a lower proportion of excellent students than in other OECD countries (Carabaña 2009), two problems that demand corrective action. One of the main weaknesses identified by PISA is the high rate of grade repetition, which triples the OECD average. Several initiatives have been put in place to address this issue, but not being able until now to solve this problem.

Consequently, although the situation clearly leaves room for improvement (albeit this is not easy), it is a long way from the disaster that some media and political leaders have sought to portray. The potential for comparison between regions has also been used for ideological and partisan purposes, even though researchers have linked existing differences with the development of the education in Spain since the nineteenth century (Martínez García 2016). Nevertheless, the impact of the Spanish participation in international large-scale studies is undeniable, as is their contribution to monitoring and improving the situation of the Spanish education system. But in general terms, it is worth noting that the general feeling is that a more intense use of this information for identifying weak points and design remedial actions is needed (Tiana 2017).

How to Go Further? Challenges and Perspectives

The current situation in the monitoring of student achievement in Spain set out above poses certain challenges that need to be adequately dealt with if significant progress is to be made in the years to come. These challenges involve a number of issues and in particular, at the least, the need to ensure the continuity of tests over

time, to facilitate the exploitation of the data obtained and to define the connection between the tests applied and individual assessment of students and schools.

As we have seen, since the beginning of the twenty-first century, changes have taken place in the approach to and nature of national tests which have prevented rigorous studies into trends and progress. Spain has infringed the rule which states 'to measure change, don't change the measure'. The consequence has been the completion of a series of cross-sectional studies, conducted at specific moments but offering little opportunity to measure change over the years.

The explanation for this discontinuity can be found in the changes that have taken place in the design and definition of the curriculum. In 2006, the constructivist curricular model of the 1990s gave way to one that placed the emphasis on key competences to be developed by students during their basic education, and in 2013, achievement standards were added that were intended to be a benchmark for assessment. These changes were undoubtedly the result of a desire to strengthen the curriculum and improve student outcomes, but they clearly introduced difficulties for maintaining the necessary continuity of the monitoring system. Here we find a general phenomenon connected to the ambivalent, two-way relationship between the curriculum defined and the assessment conducted: on the one hand, the curriculum established determines the model of assessment that should be used, whereas, on the other, the assessment system adopted (especially the approach to the tests applied) steers teaching and learning beyond the provisions of the official curriculum, since it serves as a benchmark for what is considered to be of real value in the school setting. In this case, the Spanish experience confirms the complexity of this relationship, since decisions about a specific area (like the curriculum) designed to improve outcomes can in turn have a negative impact on continuity in other areas (such as monitoring achievement).

Given this situation of discontinuity in national tests, it would seem only natural that special attention has been paid to international studies with a view to analysing the changing trends in student outcomes. Thus PISA, in particular, has become a constant reference when it comes to gauging the progress of educational performance in Spain. However, while this is undoubtedly valuable for the continuity it introduces in the evaluation mechanism, it also creates tension between the approach to and content of international tests on the one hand—which deal with certain defined skills outside the scope of national curricula—and national tests on the other, which must be based on those curricula if they are to monitor student achievement and consequently determine the internal efficacy of the system itself. In view of this ambivalent situation, it is essential to ensure the continuity of the national tests applied, with the resulting complications in agreeing on a stable curricular model. In the case of Spain, the need to reach stable agreements on the school curriculum and its application in autonomous communities and schools, despite the difficulty, is undoubtedly a prerequisite for the adoption of a stable assessment system.

In addition to the above considerations, it is worth noting that in these conditions of discontinuity, the proper exploitation of the data collected from assessment studies is no easy task, which limits any value they may have in improving outcomes.

Consequently, reports on national evaluation studies in Spain have generally been limited to the presentation of student achievement, in the form of averages and standard deviations, detailing some of the relevant aspects of the areas studied and adding, where applicable, additional correlation analysis on the influence of the socioeconomic and cultural index of students and schools. Although interesting, this provides little guidance either for teachers on how to improve lessons or to school principals on how to strengthen their modus operandi or to education authorities as to the policies that need to be reviewed or adopted. Moreover, the complexity of secondary analysis of the data provided by international studies has meant that, to date, only researchers with experience in the field have been able to exploit this constructively (Carabaña 2009, 2015; Martínez García 2016). It should be added that, as of 2013, the attempt to use national tests of an individual nature with academic consequences for students led to the replacement of the traditional multiannual sample-based survey with an annual census, making exploitation and interpretation of the data more complex in view of the increased workload imposed on assessment bodies.

All of this has had a negative effect, with the emergence of a clear imbalance between the resources mobilised for monitoring, the expectations of what it can achieve and the actual limitations of the reports produced. This problem is not exclusive to Spain, and its impact on the credibility of assessment systems has been noted on a widespread basis. The matter requires special attention if we are to strengthen the role of monitoring student achievement in the future.

On another level, there is also a need for Spain to establish what contribution it expects the system of monitoring student achievement to make, both overall and as regards individual students and their schools. The two areas should be connected, but there is no guarantee that they are. It might even be said that it is not always easy to establish this connection. This difficulty stems from whether these tests should be based on a sample- or census-based model and whether they should be annual or multiannual. Until 2013, the monitoring of individual student achievement was clearly separated from the monitoring of the performance of the education system as a whole, even though the latter was conducted by applying tests to the students themselves. Consequently, the school evaluation plans implemented were not generally based on the application of standardised tests on students but on more qualitative procedures. When the LOMCE Education Act introduced the so-called individualised assessment referred to above, the two systems became confused, as these tests (with academic consequences for individuals) also became the basis for the system of monitoring performance. This allowed the development of new school evaluation models based on the outcomes of these tests. The new model may appear coherent but has actually put considerable pressure on schools and added difficulties for assessment bodies. Schools have been tempted to establish priorities in the curriculum taught to adapt teaching to what would subsequently be evaluated (they have also done so with PISA); assessment bodies have been overwhelmed by the annual operations required by census-based assessment, preventing them from exploiting the data and drawing up detailed reports with interpretations and conclusions.

In these circumstances, we need a clear definition of the most appropriate model to conduct the different evaluation tasks (outcomes, schools and performance). There is no doubt that they may (we might even add should) be interconnected, but it is not clear that they should be part of a single system. In the specific case of Spain, this also goes hand in hand with a review of the model of the network of state and regional assessment bodies, with clear assignment of responsibilities, which means answering the question as to who is in charge of what and for what purpose. It is by no means an insignificant question. Nor is it an easy one to resolve.

In conclusion, in view of the challenges it faces, Spain needs to find an appropriate response to the issues analysed: how can it ensure the continuity of a system that enables it to conduct trend and progress analysis, how can it enable rigorous and systematic exploitation of the data obtained in a way that will help improve educational activity and its outcomes, and how can it demarcate and connect the different evaluation areas, vis-à-vis both its main stakeholders and necessary operations? The way in which Spain responds to these questions will determine the future of its system of monitoring student achievement.

References

Carabaña, J. (2009). *Una vindicación de la escuela española*. Madrid: Universidad Complutense—Facultad de Educación.

Carabaña, J. (2015). *La inutilidad de PISA para las escuelas*. Madrid: Los Libros de la Catarata.

Centro de Investigación y Documentación Educativa. (1988). *Evaluación Externa de la Reforma Experimental de las Enseñanzas Medias (I)*. Madrid: Ministerio de Educación y Ciencia—CIDE.

Centro de Investigación y Documentación Educativa. (1990). *Evaluación Externa de la Reforma Experimental de las Enseñanzas Medias (II)*. Madrid: Ministerio de Educación y Ciencia—CIDE.

Centro de Investigación y Documentación Educativa. (1992). *Evaluación Externa de la Reforma Experimental de las Enseñanzas Medias (III)*. Madrid: Ministerio de Educación y Ciencia—CIDE.

Choi, A., & Jerrim, J. (2016). The use (and misuse) of PISA in guiding policy reform: The case of Spain. *Comparative Education, 52*(2), 230–245.

Fernández-Aliseda Garrido, M. (2016). *El impacto de PISA en España: la influencia de los medios de comunicación* (Master's thesis). Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain.

Ferrer Julià, F., Massot Verdú, M., & Ferrer Esteban, G. (2006). *Percepciones y opiniones desde la comunidad educativa sobre los resultados del proyecto PISA*. Madrid: Ministerio de Educación y Ciencia—CIDE.

García Sánchez, E. (2011). *Evaluación de políticas y reformas educativas en España (1982–1992). Tres experiencias de metaevaluación*. Madrid: Instituto Nacional de Administración Pública.

Halász, G., & Michel, A. (2011). Key competences in Europe: Interpretation, policy formulation and implementation. *European Journal of Education, 46*(3), 289–306.

Instituto de Evaluación. (2009). *Evaluación General de Diagnóstico 2009. Marco de la evaluación.* Madrid: Ministerio de Educación—Instituto de Evaluación. Retrieved from http://www.mecd.gob.es/inee/dam/jcr:54064383-2ee2-4248-87a3-ce9a936ea4dd/egd-2009-marco-evaluacion.pdf

Martínez García, J. S. (2016). El PISA de hace siglo y medio. *Agenda Pública*. Retrieved from http://agendapublica.es/el-pisa-de-hace-siglo-y-medio/

Pereyra, M. A., Kotthoff, H. G. & Cowen, R., Eds. (2011). *PISA under examination. Changing knowledge, changing tests, and changing schools*. Rotterdam/Boston/Taipei: SENSE Publishers.

Roca, E. (2013). *La evaluación diagnóstica de las competencias básicas*. Madrid: Síntesis.

Tiana, A. (2014). Veinte años de políticas de evaluación general del sistema educativo en España. Revista de Evaluaciones de Programas y Políticas Públicas. *Journal of Public Programs and Policy Evaluation, 2*, 1–21.

Tiana, A. (2017). PISA in Spain: Expectations, impact and debate. *European Journal of Education, 52*(2), 184–191.

Tiana, A. (2018). Treinta años de evaluación de centros educativos en España. *Educación XXI, 21*(2), 17–36.

Tiana, A., Moya, J., & Luengo, F. (2011). Developing and implementing key competences in basic education: Analyses and reflections on curriculum design and development from the Spanish experience. *European Journal of Education, 46*(3), 307–322.

Chapter 21
Student Assessment in the Landscape of International Large-Scale Studies

Kajsa Yang Hansen and Stefan Johansson

Over the past six decades, the ILSA has changed the landscape of Swedish student assessment in many positive ways; however, it also has identified several areas of problems. In this chapter, we start with a general introduction of Swedish compulsory education system, followed by a retrospective view of the ILSA studies in Sweden, their benefits and drawbacks. This set the scene for student assessment within the framework of ILSA in Sweden today. Further, we elaborate upon the strength and weakness of Swedish grading system and national test system relating to educational quality and equity. While the national assessment systems failed to provide valid information about the trend in academic achievement, ILSA has functioned as an external support monitoring academic outcomes and educational equity over time. We present examples to illustrate the strength in the ILSA studies. Finally, we extend our perspective to the future possibilities and directions that ILSA can contribute to quality assurance in Swedish school system in general and student assessment in particular.

K. Yang Hansen (✉)
Department of Education and Special Education, University of Gothenburg, Gothenburg, Sweden

University West, Trollhättan, Sweden
e-mail: kajsa.yang-hansen@ped.gu.se

S. Johansson
Department of Education and Special Education, University of Gothenburg, Gothenburg, Sweden

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_21

Swedish Compulsory Education

Sweden has a long tradition of schooling. Although compulsory schooling was first initiated in the nineteenth century, education became systematic already in the seventeenth century when the Church Act in 1686 underlined the importance of literate citizens. Swedish was taught and tested by priests in relation to Bible reading. By the time the compulsory school system was established in 1842, more than 85% of the citizens in Sweden were literate. Children generally started school at the age of seven and continued for 6 years. This 6-year compulsory education was changed to a 9-year comprehensive school system in 1962 to meet the needs of the democratic society and rapid economic growth for citizens with knowledge and skills. In 2018 the pre-school class became mandatory, leading to 10 years of compulsory education (from age 6 to 16). Compulsory education is free of charge. Most 1–6-year-olds go to pre-school, and almost all students that attend compulsory education continue to upper secondary education (USE), albeit not being mandatory. The Swedish compulsory school system was once famous for its equity and quality. Although the educational equity and quality is still the system's major goal, deterioration has, however, been observed since a series of school reforms launched in the early 1990s. Today, the Swedish compulsory school system is characterized as one of the most decentralized education systems with school markets, deregulation of resources and free choice of school. Consequences of these reforms and remedial actions are discussed in the examples below.

International Assessment Context and Its History in Sweden

Sweden has a long involvement in student assessment in the international large-scale studies (ILSA). Since the late 1950s when IEA[1] initiated, Sweden has been an active member of the ILSA community and participated almost all the student assessment surveys conducted by IEA.

In the early phases of IEA, Sweden has had a prominent role in the organization of the IEA studies, and this was much due to the fact that Torsten Husén, one of the founders of IEA, was the chair of this organization during 1962–1978 (Husén and Postlethwaite 1996). The IEA coordinating centre was nominally located in Stockholm between 1969 and 1990, although the following chairmen worked in other locations.

Husén and his colleagues planned and launched an explorative comparative study in 12 countries, and one thousand 13-year-olds from each country were tested. The ultimate goal of this study was to investigate the methodological possibility to examine the knowledge levels of students and their determinants within and across education systems in a reliable and uniform way. This Pilot Twelve-Country Study

[1] The International Association for the Evaluation of Educational Achievement

was a success, and a couple of years later, the FIMS[2] was launched to examine mathematics knowledge of 13-year-olds and students in the final year of upper secondary school in 12 countries.

The results raised a great educational debate in Sweden, since the Swedish 13-year-olds were among the lowest in their mathematics achievement. The 1962 school reforms of the implementation of a new curriculum and an extended compulsory schooling became the target of critics for causing the low results. It is, however, not likely that reforms launched just a couple of years before FIMS would have immediate effect on students' achievement. A more reasonable explanation was that the contents of algebra and geometry tested in FIMS have not been introduced to 13-year-old students in their mathematics curriculum in Sweden (Murray and Liljefors 1983).

By many, not least media, FIMS was seen as a competition – an Olympiad. However, the initial aim of IEA was that countries could learn from each other, and the IEA studies have by no means the objective to compare and rank students' performances in all different countries. Diverse cultures, varying economies and different epistemological beliefs make it difficult to compare achievement across the range of countries (Husén 1979). Interestingly though, the 'horse-race' that was tried to be avoided for more than half a century is still an issue in ILSA today (see e.g. Klemenčič and Mirazchiyski 2018).

Already before the reports of FIMS was published, IEA began to plan a survey of six subjects, including science, reading comprehension, literature, civic education and English and French as foreign language. The results of the Six Subject Study for Sweden were corresponding to those for other Western Europe countries. The attention and debate about students' knowledge levels were far less in Sweden when the results from the Six Subject Study was released in 1973, compared to the reactions to the FIMS results.

Nowadays, IEA's mathematics and science survey TIMSS is regularly recurring with 4 years between each cycle. However, it took 16 years between FIMS and SIMS.[3] One purpose of SIMS was to investigate the changes in mathematics knowledge between 1964 and 1980. Relative to other countries, Sweden had improved its mathematics knowledge level of its 13-year-olds' population. However, the mathematics performance in SIMS was found to be at a similar level as in FIMS for students in Sweden (e.g. Murray and Liljefors 1983).

In the 1990s, and particularly with the TIMSS 95 study, a new phase was marked in the development of the ILSAs of IEA (Gustafsson 2008). This phase is characterized by a less marked researcher presence. The aim of the studies has shifted from an explanatory focus (i.e. analysing the factors behind achievement level) towards descriptive purposes (i.e. reporting descriptive outcomes; as an example, see e.g. Mullis et al. 2012). Large databases are made available for IEA studies from 1995 onwards facilitating secondary analyses.

[2] FIMS: the First International Mathematics Study

[3] SIMS: the Second International Mathematics Study

In 2000, OECD launched its Programme for International Student Assessment (PISA), which covers several subject domains, including mathematics, science and reading for 15-year-olds. In each wave, one area forms the major domain, while the other two are minor domains, represented by a smaller number of items. PISA testing is conducted every third year in all OECD countries, along with a large number of associate countries.

It should be noted that, albeit many similarities, there are some fundamental differences between the studies of IEA and OECD. PISA uses the sampling, data collection and data estimation methods and techniques similar to those used in the IEA studies. Pettersson (2008) describes how the ILSAs used by OECD emanate from those developed by IEA. However, while the IEA studies focus on curriculum-defined knowledge and skills, the OECD studies attempt to capture competencies expected to be important in adult life. Another difference concerns the definition of target population, where IEA studies sample different school grades, while PISA sample is age based. This difference means that students within the same PISA sample can come from different school grades. Thus, PISA results do not represent the learning outcomes reflecting achieved curriculum taught in certain grade. Further, while IEA may be said to have a research purpose, OECD has a policy purpose; that is, whereas IEA has developed instruments, OECD has developed a profile of policy-making (Olsen 2005).

International Assessments in Sweden Today

Today, an increasing number of ILSA studies are launched, covering different subject domains, age and aspects of education system. Sweden continues participating in IEA TIMSS, PIRLS, ICCS and OECD PISA and PIAAC[4] studies. OECD is also surveying teachers and principals in the TALIS[5] study, in which Sweden partook. As regards the performance trend in the Swedish compulsory school, Sweden has faced a general decline in the past decades. However, in the most recent assessments, an increasing performance has been observed in mathematics and science in both the TIMSS 2015 and PISA 2015. PISA 2015 even showed an improvement in reading comprehension. In TIMSS Advanced 2015, mathematical results improved, compared to 2008, but the results in physics continued to deteriorate. Albeit the recent reversed trend in achievement results, the socioeconomic gap in achievements gets continuously intensified. This indicated that the students' family background has become more important for their study results.[6]

[4] The Programme for the International Assessment of Adult Competencies (PIAAC)

[5] The Teaching and Learning International Survey (TALIS)

[6] The links to the national reports of TIMSS 2015 and PISA 2015 studies: https://www.skolverket.se/publikationer?id=3707 and https://www.skolverket.se/publikationer?id=3725

Not only the core school subjects are essential, good civic knowledge of young people is of great importance for building up a democratic society based on respect for human rights. In the International Civic and Citizenship Education Study (ICCS), Swedish eighth grade students demonstrated the highest level of knowledge of democracy, values and involvement in social issues among all participating countries. The survey also showed that Sweden is one of the countries that students significantly improved their results most compared with the previous ICCS 2009.[7]

The teacher is an important factor determining students' achievement (e.g. Hattie 2012; Gustafsson et al. 2018); therefore teachers' recruitment, retention, professional development and job satisfaction are keys to assurance of high-quality teaching and learning outcomes. OECD TALIS study offered a good opportunity for Sweden to look into, among other things, their teacher's beliefs and attitudes towards their teaching profession, the pedagogical practices and their working conditions in schools.

TALIS 2013 identified several weak points in Swedish compulsory schools. It was shown that Swedish teachers did not feel that their profession is valued in the society. Even though the great majority of teachers enjoy their work, a quite substantial amount of them would not choose to be a teacher again or regret becoming a teacher. It also indicates that it is difficult to recruit more experienced teachers to schools with higher proportion of socioeconomically disadvantaged students. In general Swedish teacher have less professional development time than TALIS average.[8]

In sum, the results from the most recent ILSA studies helped to identified several challenges in Swedish compulsory schools, which warrant policy changes and national support and investment, so that students' performance can be improved (e.g. OECD 2015).

Using ILSAs as a Tool for Monitoring Performance: an Example of International Assessment Findings in Sweden

Evaluation and assessment of educational outcomes in Swedish schools mainly are done through a combination of teachers' grading and standard national tests. The latter is used to support teachers to set school grade to their pupils. Since there is no high-stake standard testing in Swedish school system that act as recruitment criteria for school transition, the transition of school mainly relies on student's choice in combination of their school grade. Given their crucial role for teachers and schools to get the information and feedback to improve the quality of their work and for the national agency to monitor the national goals for quality and equity in education, it

[7] The link to the national report for ICCS 2016: https://www.skolverket.se/publikationer?id=3857

[8] The link to the national report for TALIS 2013: https://www.skolverket.se/publikationer?id=3293

is of primary importance that evaluation and assessment are carried out with high reliability and validity.

However, the way by which Sweden monitors its education system is not without problems. The Swedish grading system has changed a few times since the 1990s, making the comparison of how system works over time difficult. In 1994, Sweden went from a norm-referenced grading system to a criterion-referenced system. In the latter grading system, the point of reference for teachers are the goals and criteria stipulated in the syllabi. Because the criteria were multifaceted and complex in nature and sometimes abstract, and because no external point of reference was given to teachers, the new grading system suffered from great variation in teachers' grading practice across schools and over time. One of the consequences of such a variation is grade inflation, particularly so during the first years after the introduction of the new grading system in 1994, which highly affected the fairness in assessing students' school performance (i.e. Cliffordson 2004; Vlachos 2010). In the 2011 curricula, criteria have been revised again, and the grading system has changed from a four-point scale to a six-point scale with somewhat more detailed knowledge demands for students' knowledge and skills. It is, however, little evidence that the new curricula and syllabi would improve the equality and equity in grading (Gustafsson et al. 2014).

The national tests are administrated in Sweden in several grades and school subjects; however, the characteristics of the tests have varied greatly across the different administrations, and most often, teachers correct their own students' tests. As Gustafsson et al. (2014) noted, teachers' assessments of most of the subjects in the national tests are more lenient than those of the external markers. Another problem is that there is large variation in test scores from 1 year to another. Moreover, the proportion of students who do not pass the national test of mathematics, for example, is clearly higher than the proportion of students who failed in their final grade. These observations indicate that the national test scores, at least in certain subjects, fail to support teachers' grading practice, nor can they be used as a foundation for analysis about the degree to which the national knowledge requirements are fulfilled in different school types and across the nation as a whole (Gustafsson et al. 2014). These weaknesses in assessment practices affect the validity and reliability of the intended function of school grades and the national tests in monitoring quality and equity goals in the Swedish school system.

As is shown in Fig. 21.1, the average school grade has increased over the last 30 years; however, an opposite trend has been observed by PISA data. In Sweden, declining achievement and equity has been noted since the early 2000s. Hanushek et al. (2012) analysed the ILSA achievement trend in 49 countries and found Sweden being the country with the most severe decline. Gustafsson et al. (2016) describe that the general knowledge of university students has been reported to decrease in the past decades. They argue, furthermore, that the teacher education lost status, in that salaries are low, and because the profession no longer attracts high-performing students.

One hypothesis is that the decline is related to the decentralization, deregulation and marketization reforms in the late 1980s and early 1990s. These reforms have

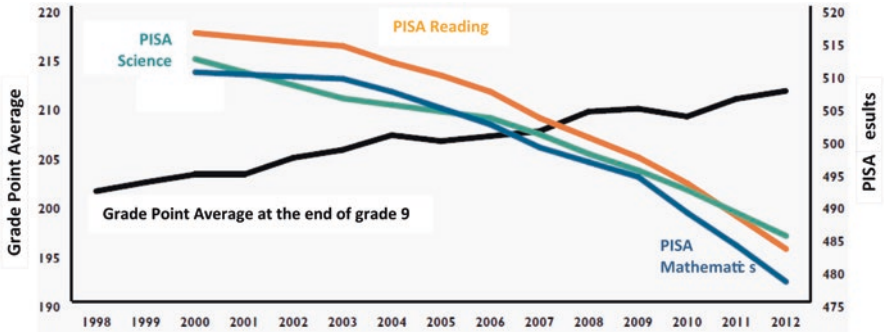

Fig. 21.1 Different trend between school grade point average and results from PISA studies. (Source: Henrekson and Jävervall 2016)

transformed the Swedish educational system in ways that may have led to changes in students' learning opportunities. As the marketization of the Swedish school system expanded rapidly in the first decade of the new millennium, the number of private-run and public-funded for-profit independent schools increased largely, and the school system has become a more segregated one. Meanwhile, Sweden also faced a substantial immigration, deteriorating social welfare and increasing income gap. However, the increasing immigration alone cannot lead to a decline in achievement according to previous research (Skolverket 2012). One possible cause of the achievement decline in Sweden has to do with factors at system level. In a school system with strengthened between-school socioeconomic and ethnic segregation due to different kinds of selection (e.g. tracking) or self-selection (e.g. free choice of school) mechanisms, the variation in socioeconomic and ethnic composition of school intakes increase across schools. Since family background is one of the influential factors for student's academic achievement (Sirin 2005), the school social and ethnic segregation leads to a different learning environment, peer dynamics, pedagogical resources and instructional quantity and quality. These factors in turn exacerbate educational inequality and reduce educational quality.

Studies evaluating effects of these reforms are, however, so far limited and inconclusive (Holmlund et al. 2014). This may partially be due to methodological challenges, due to lack of appropriate data. However, ILSA data outcomes provide a lens through which Sweden can examine its own practices and policies in education. Especially, when the effects of some institutional characteristics are impossible to study due to lack of variation within a single country (e.g. the existence of national exams), comparative studies are the only approach to observe variation in these features. Even though differences in student achievement typically attract attention in these studies, the true value of international comparisons lies in the consideration of policies related to the structure; organization, delivery and content of instruction as a range of alternatives are brought to light, allowing us to examine achievement trends and their determinants in Sweden, using other educational systems as points of reference.

Therefore, external assessments, such as ILSA studies conducted by IEA and OECD, can offer further accountability measures to examine their school quality and equity goals in Sweden (see, i.e. Nusche et al. 2011). The following is an example that demonstrates the use of ILSA data to identify factors that can compensate disadvantages in family background and improve educational equity and quality in Sweden and around the world.

Using data from the eighth graders in 50 countries participating in the TIMSS 2011, Gustafsson et al. (2018) examined the influence of quality and quantity of teaching, school climate and school SES composition on the relationship between individual student's family socioeconomic status (SES) and their mathematics achievement. The relationship between SES and mathematics achievement is used as a measure for educational equity. A high relationship implies that student's family SES has high effect on their academic achievement, thus indicating a low educational equity.

The study applied the so-called random slope model to capture the variation in the relationship between student's SES and achievement (i.e. educational equity) across different schools within each education system. The random slope was later on regressed on different school characteristics, e.g. instructional quantity, instructional quality, school climate, school emphasis on academic success and school SES composition. The study found that school characteristics reflecting quality and quantity of instruction, school climate and school SES are potential determinants of educational equity across school systems. However, these factors may function differently for countries in different state of development. In highly developed countries, these factors may compensate the effect of students' family SES on their academic achievement, and students of lower-SES families may benefit particularly and thus help to improve educational equity. Sweden is one of these countries. Highly developed countries are well-advised to pay more attention to school characteristics if they intend to make their school system a more egalitarian one and to promote educational equity.

For developing countries, these school characteristic factors do not necessarily help to improve equity. Since most of these school systems may be elitist educational systems, students from higher SES families profit strongly from high instructional quality, academic success emphasis by school, and an orderly climate. This may also indicate, for example, that there are generally better teachers and resources in schools with students from high-SES families. Such education systems may cause socioeconomic segregation. The study also found that the cross-level interaction between the school SES composition and the within-school SES – achievement relationship – is a powerful indicator of a school system's educational equity. And educational equity tends to go hand in hand with educational quality.

The interpretation of these results is related to the organizational differentiations across different education systems. In Sweden, for example, since the school reforms in the past decades, especially the launch of free choice of schools with a nationwide voucher system and independent school reform, the Swedish compulsory schools become more segregated with respect to school composition of SES and ethnicity of their students. This interpretation has been confirmed in other

studies based on other ILSA data and national data (e.g. Yang Hansen et al. 2014; Yang Hansen and Gustafsson 2016). Moreover, teaching quality is considered to be a crucial school-level factor predicting student academic outcomes (e.g. Hanushek et al. 2014). The decentralization reform, however, gave schools and municipalities more autonomy in allocating resources in education and recruiting teachers. Also, good teachers tend to select schools with students from families of the same socio-economic background as themselves, which have caused an increasing diversity among schools with respect to educational resources and teaching quality (e.g. Han 2018; Holmlund et al. 2014; Skolverket 2009). Schools and teachers have high degree of autonomy to decide teaching contents, methods and materials. Schools also offer different study options, and the interpretation of the goal documents varies largely among teachers. These have a strong impact on equal education opportunity and in tune lead to the achievement disparity.

Discussion and Conclusions

The examples presented in previous sections have shed light on how ILSAs have been utilized in recent past. ILSA studies do have a prominent position in educational research with its increased popularity in terms of number of studies and public attention. The results of ILSA's in Sweden have, as in many other countries, raised concerns about the quality, equity and efficiency of Swedish school system. The results in PISA 2012, for example, caused a collective shock in Sweden. A decrease in performance in all three subjects in PISA 2012 – mathematics, science and reading comprehension – was observed, and achievement levels were all below the OECD average. Sweden was also the country that deteriorated the most in this PISA study, making the school issue one of the most important issues for the 2014 elections in Sweden. A national debate on how to raise the quality of school education and to build a broad consensus on changes in the education system was thus fostered. Based on the research evidences, challenges in Swedish school system were identified, among other things, inadequate capacity building and conditions for professional development for teachers, low teaching profession status, weaknesses in resource allocation, school segregation and quality diversity across schools, problems with the learning environment and incoherence and unreliable assessment and evaluation (OECD 2015; Statens Offentliga 2017).

To respond to these weaknesses and to improve educational outcomes and equity, Swedish school commission proposed a series of policy suggestions for education system amendment, as further reinforcement to some earlier actions implemented successively. For example, to raise competence for teachers and school leaders, the national agency of education started boost for teachers (lärarlyftet) in 2007 and boost for mathematics (matematiklyftet) in 2012, and the National School Leadership Training Programme was made compulsory in 2011. In the same year, a new curriculum and syllabi for compulsory school were passed with clearly and concretely stated learning standards and goals. The new syllabi also include more

description of teaching content to ensure that learning opportunity for every child in Sweden dwells on a common basis of knowledge and skills (Skolverket 2011). Meanwhile the idea of evidence-based policy-making was put forward through strengthened emphases on school assessment and evaluation and accountability to generate and analyse high-quality data from different levels. A very similar example also has been observed in Germany (Grek 2009).

After the PISA shock in 2012 in Sweden, the school commission suggested to continue with the on-job capacity building for teachers and school leaders. In addition, further actions have been taken to, for example, optimize resource allocation according to socioeconomic and ethnic composition of school intakes, so that schools with large proportion of children from disadvantaged families can get extra supports needed from schools and teachers. Most importantly, with special educational efforts on preventing difficulties of various kinds and giving each student the opportunity for optimal learning, a good social climate is warranted. Recent investment of a total of 540 million SEK in educating more teacher with special pedagogical orientation (specialpedagogiska lyft) in Sweden showed the attempt to improve school performance, equity and inclusion for all.

In sum, these actions taken in Swedish education system recently demonstrated the connection between the school quality and equity and educational reforms and policy changes. The empirical evidences from these studies thus raised cautions and debates concerning amendment actions that Swedish education system needs to be taken to reverse the deteriorating trend in school quality and equity. Thus, ILSA studies have a prominent position in 'evaluation and policy discussions within participating countries' and act 'as an infrastructure for research' (p. 329, Gustafsson 2018).

Although there are indications that ILSA has impact on school systems, which sometimes may be unintended, the long-term effects are difficult to trace and causally relate to ILSA results. It should be remembered that borrowing and travelling ideas of educational policies are mechanisms of change that are subtle and difficult to chart and analyse, not least in terms of long-term effects. One reason for this is that what has been written on paper does not imply change in actual teaching or achievement (e.g. Steiner-Khamsi and Stolpe 2004).

Finally, this chapter may be viewed as a contribution to the discussion about the credibility of ILSAs. In contrast to much research on ILSAs, we would like to point out that there are benefits associated with the production of data generated from the numerous studies. We believe that researchers will continue to utilize ILSA data for a long time. Data has a longitudinal component at the country level, which facilitates opportunities to investigate causal effects of the impact of different reforms in different countries. The methodological advances in this field have been substantial in recent decades. Few infrastructures in the world cover such a lengthy time span or such a large number of participants.

References

Cliffordson, C. (2004). Betygsinflation i de målrelaterade gymnasiebetygen [Grade inflation in the goal-oriented upper secondary school grades]. *Pedagogisk Forskning i Sverige, 9*(1), 1–14.

Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy, 24*(1), 23–37. https://doi.org/10.1080/02680930802412669.

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal, 7*(1), 1–17.

Gustafsson, J.-E. (2018). International large scale assessments: Current status and ways forward. *Scandinavian Journal of Educational Research, 62*(3), 328–332. https://doi.org/10.1080/00313831.2018.1443573.

Gustafsson, J-E., Cliffordson, C., & Erickson, G. (2014). *Likvärdig kunskapsbedömning i och av den svenska skolan – problem och möjligheter*. [Equity in assessments on knowledge and skills in Swedish school: Problems and possibilities]. Stockholm: SNS Förlag.

Gustafsson, J.-E., Sörlin, S., & Vlachos, J. (2016). *Policyidéer för svensk skola* [Policy ideas for Swedish Schools, in Swedish]. Stockholm: SNS Förlag.

Gustafsson, J.-E., Nilsen, T., & Yang Hansen, K. (2018). School characteristics moderating the relation between student socio-economic status and mathematics achievement in grade 8. Evidence from 50 countries in TIMSS 2011. *Studies in Educational Evaluation, 57*(special issue), 16–30. https://doi.org/10.1016/j.stueduc.2016.09.004.

Han, S. W. (2018). School-based teacher hiring and achievement inequality: A comparative perspective. *International Journal of Educational Development, 61*(July), 82–91. https://doi.org/10.1016/j.ijedudev.2017.12.004.

Hanushek, E. A., Peterson, P. E., & Woessmann, L. (2012). Achievement growth: International and U.S. State trends in student performance. Retrieved from http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%2BPeterson%2BWoessmann%202012%20PEPG.pdf

Hanushek, E. A., Piopiunik, M., & Wiederhold, S. (2014). *The value of smarter teachers: International evidence on teacher cognitive skills and student performance* (NEBR Working Paper Series). Cambridge, MA: National Bureau of Economic Research.

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. London: Routledge.

Henrekson, M. & Jävervall, S. (2016). *Svenska skolresultat rasar – vad vet vi?* [Swedish school results falling – What do we know?]. Stockholm: The Royal Swedish Academy of Engineering Sciences (IVA). Retrieved from https://www.iva.se/globalassets/info-trycksaker/iva/201609-iva-henrekson-javervall-i.pdf

Holmlund, H., Häggblom, J., Lindahl, E., Martinson, S., Sjögren, A., Vikman, U., & Öckert, B. (2014). *Decentralisering, skolval och fristående skolor: Resultat och likvärdighet i svensk skola* [Decentralisation, school choice and independent schools: Results and equity in Swedish schools] (IFAU Report No. 2014: 25). Retrieved from http://www.ifau.se/globalassets/pdf/se/2014/r-2014-25-decentralisering-skolval-och-friskolor.pdf

Husén, T. (1979). An international research venture in retrospect: The IEA surveys. *Comparative Education Review, 23*(3), 371–385

Husén, T., & Postlethwaite, T. N. (1996). A brief history of the international association for the evaluation of educational achievement. *Assessment in Education: Principles, Policy & Practice, 3*(2), 129–141. https://doi.org/10.1080/0969594960030202.

Klemenčič, E., & Mirazchiyski, P. V. (2018). League tables in educational evidence-based policy-making: Can we stop the horse race, please? *Comparative Education, 54*(3), 309–324. https://doi.org/10.1080/03050068.2017.1383082.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Murray, Å. & Liljefors, R. (1983). *Matematik i svensk skola* [Mathematics in Swedish schools, in Swedish]. Utbildningsforskning, FoU-rapport 46. Stockholm: Skolöverstyrelsen och Liber Utbildningsförlaget.

Nusche, D., Halász, G., Looney, J., Santiago, P., & Shewbridge, C. (2011). *OECD reviews of evaluation and assessment in education: Sweden 2011*. Paris: OECD.

OECD. (2015). *Improving schools in Sweden: An OECD perspective*. Paris: Organization for Co-operation and Development.

Olsen, R. V. (2005). *Achievement tests from an item perspective. An exploration of single item data from PISA and TIMSS studies, and how such data can inform us about students' knowledge and thinking in science*. (Doctoral dissertation). Retrieved from https://www.duo.uio.no/handle/10852/32287

Pettersson, D. (2008). *Internationell kunskapsbedömning som inslag i nationell styrning av skolan.* [International knowledge assessments: an element of national educational steering] Acta Universitatis Upsaliensis. Uppsala Studies in Education No 120. Uppsala: Uppsala universitet

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453. https://doi.org/10.3102/00346543075003417.

Skolverket. (2009). *Vad påverkar resultaten i svensk grundskola? Kunskapsöversikt om betydelsen av olika faktorer* [What influence the achievement results in the Swedish compulsory school?]. Stockholm: Skolverket.

Skolverket. (2011). *Läroplan för grundskolan, förskoleklassen och fritidshemmet 2011*. Stockholm: Skolverket. http://natprov.edu.uu.se/digitalAssets/173/c_173998-l_3-k_lgr11.pdf

Skolverket. (2012). *Likvärdig utbildning i svensk grundskola? En kvantitativ analys av likvärdighet över tid* [Equal education in Swedish elementary school? A quantitative analysis of the equality over time]. Stockholm: Skolverket.

Statens Offentliga Utredningar. (2017). *Samling för skolan Nationell strategi för kunskap och likvärdighet, Slutbetänkande av 2015 års skolkommission* [Collection for school national strategy for knowledge and equivalence]. Stockholm: Staten Offentliga Utredningar.

Steiner-Khamsi, G., & Stolpe, I. (2004). Decentralization and recentralization reform in Mongolia: Tracing the swing of the pendulum. *Comparative Education, 40*(1), 29–53. https://doi.org/10.1080/0305006042000184872.

Vlachos, J. (2010). *Betygens värde – En analys av hur konkurrens påverkar betygssättningen vid svenska skolor* [Value of the school grade – An analysis of how competition affects grades at Swedish schools, in Swedish]. Stockholm: Swedish Competition Authority.

Yang Hansen, K., & Gustafsson, J.-E. (2016). Causes of educational segregation in Sweden – school choice or residential segregation. *Educational Research and Evaluation, 22*(1–2), 23–44.

Yang Hansen, K., Gustafsson, J.-E., & Rosén, M. (2014). School performance differences and policy variations in Finland, Norway and Sweden. In A. B. Kavli & T. Hallvard (Eds.), *Northern Lights on TIMSS and PIRLS 2011: Differences and similarities in the Nordic countries* (pp. 25–48). Nordon: Nordic Council of Ministers. https://www.udir.no/Upload/Forskning/2014/Nlights%20TIMSS%20and%20PIRLS.pdf

Chapter 22
European Monitoring of Student Achievement in the Twenty-First Century: Summary and Outlook

Justine Stang, Nele McElvany, and Heidi Harju-Luukkainen

There is a large body of research literature pointing out to the benefits of education for the individual as well as for society. Education can therefore effectively level the societal playfield and improve the quality of life as well as increase opportunities for individuals in a long-term perspective. International assessments and national monitoring instruments have a crucial role here. They provide relevant information on the learning conditions and educational outcomes for the policymakers and practitioners. This information can then be used for curriculum reforms, improvement of the overall educational systems, promoting educational equity or, for instance, towards improving the teaching and learning processes. The high importance and necessity of international large-scale assessments are reflected in two main goals pursued. One main goal is the monitoring of student achievement. The second main goal is benchmarking. As a result of this, international large-scale assessments can be used to draw comparisons with educational standards. This means that large-scale assessments are used to systematically and regularly examine the quality of educational systems. Data of large-scale assessments provide information on students' competences in different domains and hint on different process and context factors, which shed light on how the individual educational system works. Additionally, due to its internationality, the results are used to draw comparisons with the results of other countries and their educational systems.

The origins of our present international large-scale assessments of different student skills are based on the First International Mathematics Study (FIMS), more than 50 years ago. Since then, the large-scale assessment landscape developed immensely. The development over the last 50 years and the importance of large-

J. Stang (✉) · N. McElvany
Center for Research on Education and School Development (IFS), TU Dortmund University, Dortmund, Germany
e-mail: justine.stang@tu-dortmund.de

H. Harju-Luukkainen
Nord University, Levanger, Norway

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_22

scale assessments raise the question, how educational assessment looks like in the twenty-first century of Europe. Therefore, the book focuses on the one hand on general aspects of educational assessment like aims and approaches of monitoring, history and current state of large-scale assessments as well as on methodological challenges, complemented by an American perspective on student monitoring. On the other hand, the book focuses on several European countries in order to give an overview of their assessment history, their current state as well as on how results are used and whether they do have an effect, for example, on changes in curriculum. Due to that, differences and commonalities between these European countries can be detected. Therefore, this chapter sums up briefly some differences and commonalities between these European countries in relation to the history of international large-scale assessments, the present situation and their performances in international large-scale assessments in order to focus on how results of large-scale assessments are used.

First of all, the European countries in this book differ in their participation history in international large-scale assessments. In FIMS (1964) as the first large-scale assessment, the European countries Belgium, Finland, Germany and Sweden took part. In contrast, the large-scale assessment history of other European countries in this book is shorter. Nevertheless, it is important to note that it is in common for almost all of these countries that the participation in international large-scale assessments was not continuous, which means that many European countries had large gaps in their participation rate in large-scale assessments.

In contrast to the beginning of large-scale assessments where only a handful of countries took part, international large-scale student assessments are nowadays a constant and relevant element for all of the European countries in this book. For example, all these European countries took part in the known and relevant assessment Programme for International Student Assessment (PISA; 2015), while not all of them – but most – participated in the last cycle of the Progress in International Reading Literacy Study (PIRLS) or the Trends in International Mathematics and Science Study (TIMSS). All in all, in relation to the participation in international large-scale assessments, the European countries are well positioned. As opposed to this, national large-scale assessments are not so far advanced in every European country in this book but are becoming more and more important as well – also due to results of international large-scale assessments that demonstrated the necessity and importance of student monitoring on a national level.

Focusing on international large-scale assessments that are generally known in public, we can highlight differences and commonalities in the performance levels of the different European countries in this book. The countries differ not only in the performance level concerning above, below or at the average, but also in the development over the years. This means that the patterns of development are different. Some countries present in this book improved, some stayed stable, and some achieved lower average scores. The differences in performance levels and in developments are similar for several international large-scale assessments. Furthermore, the reasons which are named for differences in performance levels and in the development differ between the countries. For example, for countries that scored above

the average, the results typically were interpreted as confirming that the education strategy worked well, while for some countries that scored below the average, the results were often seen as a kind of critique and led to doubts of the efficiency of the educational systems. Differences in performance levels, especially for countries scoring below the average, and negative patterns in development led to different types of discussions and to stronger or weaker effects on changes in curriculum, assessment strategies or support for subgroups of students with particular low average achievement levels.

The results of international large-scale assessments illustrated either strengths or weaknesses of educational systems reflected by the performance ranking. As mentioned before, this had an impact on how results of large-scale assessments were used. Therefore, differences and commonalities exist in relation to the effects on changes, for instance, in curriculum and usage of results. Common for all these European countries is that the results of international large-scale assessments provide feedback at several levels which is relevant for and is perceived by teachers, schools, administrators and education policymakers. Additionally, it is common for many European countries that results of international large-scale assessments (LSA) led to an increased interest in outcomes of educational systems as well as on educational research itself. Moreover, LSA had an impact on empirical educational research themes. Important topics that were analyzed are, for example, tracking systems, instructional quality, teacher education, social inequalities or migration background. Another important feature of LSA was utilized by European countries in this book as well: Results are being used to compare their own performance with other significant countries. Furthermore, the European countries in this book as well as other countries that took part in international large-scale assessments face the problem of interpretability of the data due to its cross-sectional design. Last but not least, many European countries face the challenge of shifting from paper-and-pencil to computer-based assessments which goes hand in hand with methodological issues and usage of results.

Opposed to these commonalities, also many differences exist in relation to effects and usage of international large-scale assessment results across European countries. The effects on changes, for instance, in curriculum, strategies or educational systems are multifaceted. Depending on the performance ranking, results of international large-scale assessments uncovered different problems that were of political, academic or public interest. Consequently, results led to stimulated debates on, for example, the quality of educational systems and on problems of the educational systems such as teacher education or social inequality, whereas European countries in this book focused differently strong on these topics. For those European countries that perceived so-called shocks (e.g. PISA-shock), it is in common and therefore contrary to those that did not perceive ‘shocks’, that they led, for instance, to the implementation of educational standards, to stronger reforms of educational systems, to additional support possibilities for teachers or to more structural changes in curriculum as well as to a stronger promotion of national evaluation systems. This implies that lower-performing countries have in common that results of international large-scale assessments led to an increased awareness of the need for changes

concerning different aspects. The awareness of problems, the scrutinizing of possible factors and addressing these issues in research led also to a stronger acceptance of international large-scale assessments and to – at least for some countries – better results.

This book has highlighted through its 22 chapters different monitoring policies and practices across Europe. It has given readers an opportunity to understand the different strategies and used indicators. When it comes to monitoring and developing policies and practices, there are many similarities across the chapters of this book. One way of measuring these is the repetitiveness of words used across the chapters. One hundred of the most commonly used words of this book are presented in Fig. 22.1. The size of the word in the figure indicates its frequency in book (ranging from 85 to 1015). What this figure clearly highlights is that the topics of discussion across the chapters are strongly related to OECD or IEA assessments, especially to PISA, PIRLS and TIMSS. These can therefore be seen as the most dominant international assessments connected to policies and practices across the different countries presented in this book. The high weight of IEA and OECD led assessments on educational policy and practice is calling for attention to a deeper discussion on the role of international and national assessments in every country.

Policymakers and researchers are keen to learn from international comparisons and develop policies and practices based on them. However, there are still many things that are yet to be taken into consideration when conducting and especially interpreting the results of international assessments across different countries. It is important to note that comparing assessments like PISA or TIMSS across different cultures is not problem free. This is something that can be observed across the

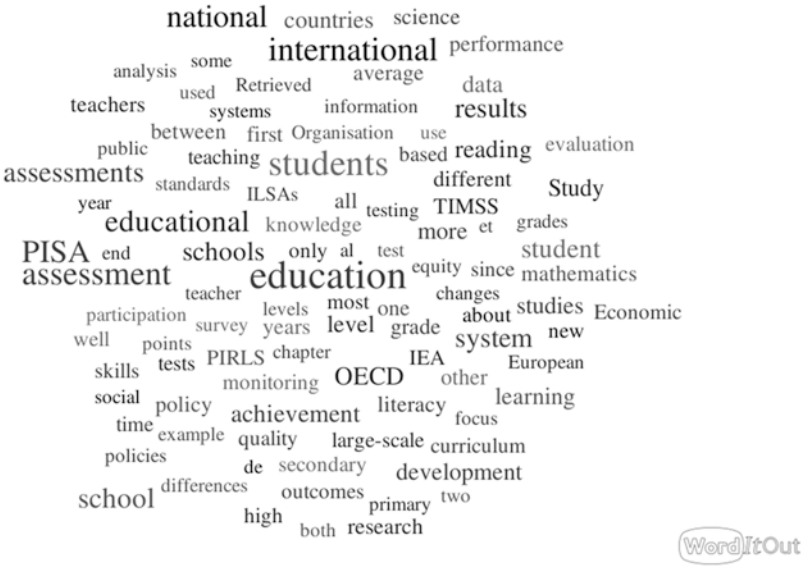

Fig. 22.1 Wordle cloud indicating 100 mostly used words across the chapters

different chapters of this book. According to He et al. (2019), for both cognitive and non-cognitive assessments, the presence of bias (construct, methods and item) indicates that scores from the assessment in different cultures reflect some cultural characteristics other than what the assessment is intended to measure. Therefore, before any comparative inference is made, biases need to be detected and ruled out. As a result of this, measurement invariance testing in order to check the comparability is of high importance (see closer He et al. 2019). This all is essential to keep in mind before coming up to robust conclusions that might have long-term effects on the policies and practices of the education system.

Taken together, international large-scale assessments are an important tool in order to monitor or for benchmarking student achievement, but as well they have limitations that need to be understood and discussed. These limitations need to be understood not only by researchers but also by policymakers. This is because the results of national and international assessment can be very powerful as they cause stimulating debates on the quality of educational systems going hand in hand with a critical reflection of existing challenges. Based on empirical evidence, possible measures of actions in order to tackle these diverse issues can be identified. Living in an increasingly globalized world, it is important to know and to understand the assessment context, strategies and results of other countries as well as similarities and differences of assessments in order to understand, reflect and reassess critically own perspectives, strategies and results.

Reference

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice, 26*(4), 369–385. https://doi.org/10.1080/0969594X.2018.1469467.

Correction to: Assessment Policy and Practice of Slovenia

Klaudija Šterman Ivančič and Urška Štremfel

Correction to:
Chapter 19 in: H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_19

The original website of the link that was mentioned in the chapter 19 reference: *Šterman Ivančič, K., & Puklek Levpušček, M. (2018) Individual and teacher-level predictors of student achievement: PISA 2015. Presentation from EARA 2018. Ghent*: *European Association for Research on Adolescence* has been hacked and removed as it leads to an inappropriate website.

Therefore, the link has been removed and corrected reference is listed in chapter 19 as follows:

Šterman Ivančič, K., & Puklek Levpušček, M. (2018) Individual and teacher-level predictors of student achievement: PISA 2015. Presentation from EARA 2018. Ghent: European Association for Research on Adolescence.

The updated online version of the chapter can be found at https://doi.org/10.1007/978-3-030-38969-7_19

H. Harju-Luukkainen et al. (eds.), *Monitoring Student Achievement in the 21st Century*, https://doi.org/10.1007/978-3-030-38969-7_23