Large-scale assessments in the Norwegian context

Henrik Galligani Ræder, Rolf Vegar Olsen and Sigrid Blömeke

BIO

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

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

Sigrid Blömeke is director and professor at the Centre for Educational Measurement at the University in 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.

Abstract

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

The foundation of education in Norway is a comprehensive school system during the first 10 years of schooling followed by a three-year upper-secondary program. This chapter discusses the large-scale assessments in place during the first 10 years of the Norwegian school system. Education at the primary (year 1-7) and lower secondary (year 8-10) levels is predominantly public and free, and until selection into upper-secondary school, any room for school choice is very limited (Organisation for Economic Cooperation and Development, 2013).

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 decentralized 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 is 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, Hopfenbeck, Ibsen, & Turmo, 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, Prøitz, & Skedsmo, 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 five 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. wellbeing, 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 grade 5 to 13. Approximately 75% of students in grades 5–13 participated in the most recent surveys (Wendelborg, Røe, Utvær, & Caspersen, 2017).

ILSAs in Norway

Since 1995, Norway has participated in almost all cycles of the major ILSAs organized 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 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 one year of schooling in both the 4th and the 8th 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 & Lagerstrøm, 2014).

<FIGURE 1 HERE>

In the last half of the period represented in Figure 1, the trend is reversed with an approximately equally large increase in scores across cohorts and domains (with 8th 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, Björnsson, & Olsen, 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, Björnsson, & Olsen, 2018). Finally, it should be noted that the decline in Norway's achievement from 1995 to around 2003/06 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 & 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 & 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 standardized 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 jeopardizing 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 & 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 & 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, Kovac, Kvåle, Langfeldt, & Skov, 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 is 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 grade 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 makes linking hard, if not impossible. The mapping tests are optimised to have maximum information at the cut-score (20^{th} 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 grade 5–9. 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 standardized, casting some doubt about the comparability of the data across schools (Wendelborg & Caspersen, 2016). From a researchers point of view, efforts should be made to rectify these potential sources for bias at the school level. However, this should not come at the cost of the current advantages.

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 & 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 & 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).

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. doi: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 (2013). *Education policy outlook 2013: 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). Münster: Waxmann.

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.

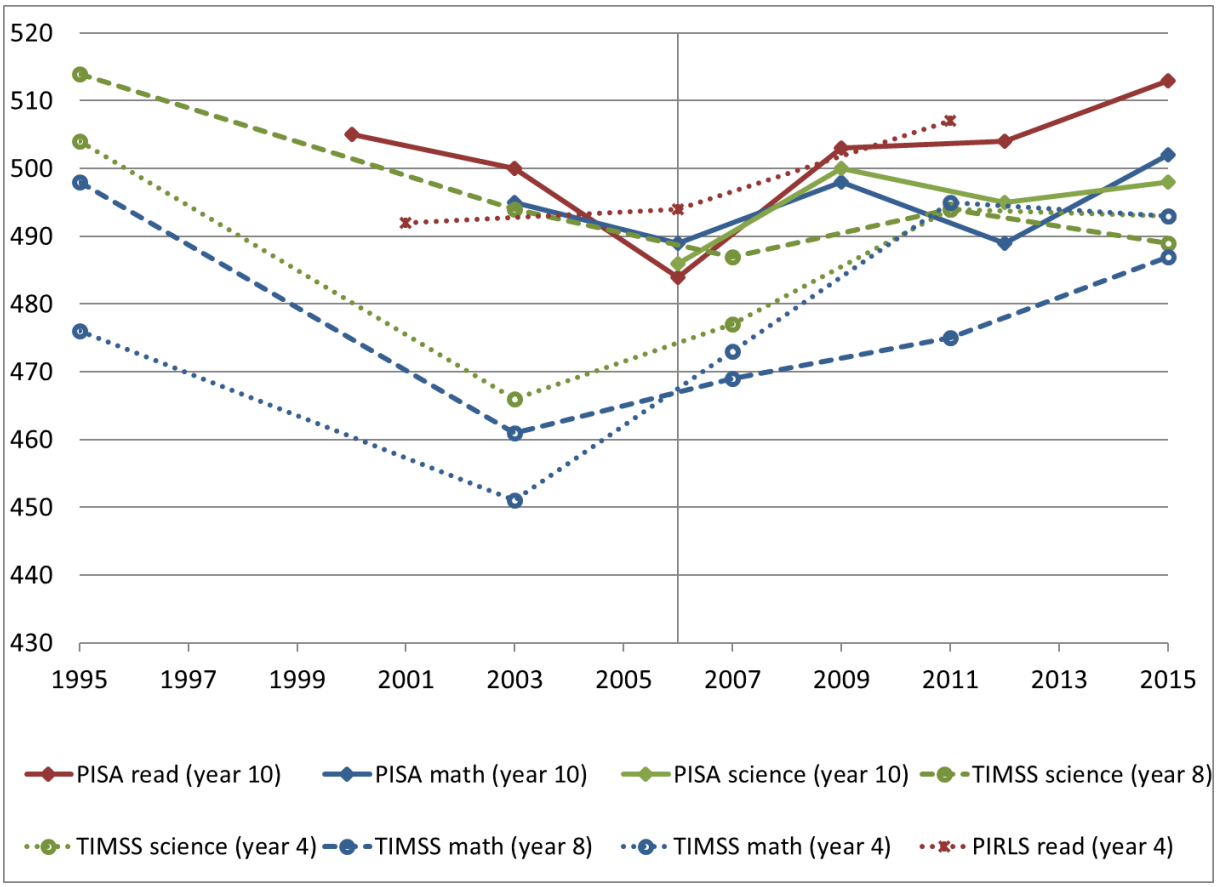

Figure 1. Average scores for all populations in PISA, PIRLS and TIMSS 1995 to 2015. Scores for each study are represented by the originally reported scales from the studies (taken from Olsen & Björnsson, 2018, p. 21).