UiO **Centre for Educational Measurement (CEMO)**
Faculty of Educational Sciences

An evolving framework for CEMO's research relating to the Norwegian context (Version 1)

Contents

This is a dynamic document. We are grateful for any reactions, thoughts and comments you may give us

1. Background

The strategic documents for CEMO describe a wide range of research problems in the field of educational measurement of relevance for the activities at the centre. At the one end, the field is concerned with understanding measurement in its fundamental sense. This line of research may be labelled as fundamental methodological research and is at the heart of CEMO's mandate. At the other end of the range, insights from fundamental research of measurement and other methodological issues are brought into more applied research settings where measurement is a central part of the design, but where the research questions are substantially about education and learning. CEMO also has a mandate to support research and development of measurement-intensive applied research areas in education – from kindergarten to higher education. In particular, CEMO seeks to support research on testing and measurement in the practical school context as an object or phenomenon in itself[1].

The background for the establishment of CEMO was a perceived need for strengthening the research capacity in Norway in the field of educational measurement. CEMO has been successful in recruiting internationally very strong scholars in the field. The centre has also been successful in establishing external research funding. In particular, the centre has been able to establish a strong research portfolio of methodologically oriented project.

Relating more specifically to the national context, CEMO has substantial research activities within higher education and some activities relating to early education and child care. However, the centre has so far not established much research activity explicitly targeting assessments in primary and secondary school in Norway. This document presents what should be perceived as a dynamic and evolving framework to establish and nurture future activities relating to the latter context at the centre.

The document contains two main parts. The first part describes the context of assessment in the Norwegian primary and secondary schools, the intention of which is both to help identify research potential and to introduce vital aspects of the context to those not familiar with the educational system of the country. We are also pointing to a variety of different relevant sources, both academic and non-academic material. Where possible, we have given the link to freely accessible sources, but some references are also given to academic papers in journals.

The second part of the document identifies and provides brief descriptions of a selection of specific topics or areas where research involving the national context is needed. As this is a dynamic document, the ambition is that some of the areas or topics in this part should be further developed at a later stage into complete research proposal. Initially one or two of these research projects will be established through internal funding of PhD positions. In the somewhat longer perspective the ambition is that one or more of these topics is further developed into research proposals submitted to the National Research Council (NFR).

[1] See more at http://www.uv.uio.no/cemo/english/research/mission-statement/

2. Relevant features of the Norwegian context[2]

The dominant element of assessment in the Norwegian system is a large trust in teachers' own judgements of students. In practical terms this means that teachers are given a mandate for both formative and summative assessment purposes[3]. Basically, the Norwegian system for grading is criterion-based where teachers' are given the task to assess the degree to which their pupils demonstrate mastery of the aims stated in the national common curriculum. The nature of the curricular aims varies across subject domains, but they are all (more or less) expressed as competencies, or in other words, as expressions of what students should be able to do at certain infliction points in the ten year compulsory education (after grade 2, 4, 7 and 10). Similar curriculum documents also exist for upper secondary education.

Teachers' grading decisions are based on monitoring and observing pupils over time in the classroom, and by grading of single performances of various kinds (presentations in the classroom, written essays, teacher-made tests etc.). A relatively recent shift of the legislation has changed how teachers should relate to such assessment products over the school years. It is now clearly expressed that the final grading in the subject should not be an aggregate of these single observations over the school-year, but rather should reflect the mastery by the end of the school-year. In general little knowledge exist regarding the forms of assessments applied by teachers, and how the information gained from the various assessment occasions over the school year are used in the final grading decisions.

In addition to curricula for specific subjects, there are also separate frameworks for five cross-curricular domains, labelled as basic competencies in reading, writing, numeracy, ICT and oral skills. These are not subjects taught in school and pupils are not explicitly graded in these domains. Instead, the idea is to integrate the cross-curricular domains into the subjects. These domains are of specific interest in the context of assessment in Norway. A majority of the nationally developed low-stakes tests which have been introduced in Norway over the last decade relate to these basic competencies.

Students are kept in comprehensive schools and classrooms until grade 10 (primary grades 1-7, lower secondary grades 8-10). School intake is in general based on residency in an area[4], and there are only a small percentage of students attending private schools[5]. As a consequence, the amount of variance in achievement associated with the classroom and school level is very small in Norway[6]. Grade 11 is the first year in upper secondary school. This is also the first year that students are streamed and tracked into

[2] For more detailed descriptions of the Norwegian educational system in English, see https://www.udir.no/in-english/

[3] For broader perspectives on the Norwegian assessment system see Nusche, Earl, Maxwell, & Shewbridge (2011) and Tveit (2014).

[4] Some larger municipalities, with Oslo as the prime example, have installed a policy of free school choice. However, every kid is granted intake to their locale school, and the schools will only accept external applicants if there are places available in the already existing classrooms.

[5] The private schools are either based on religious or alternative pedagogical ideologies, and they are predominantly publicly funded.

[6] Figures in the interval 5-10 % are typically reported. Details may be found in the reports from international studies such as TIMSS and PISA

different educational programs. There are five programs leading to a general academic qualification, and eight programs leading to a large number of specific vocational certifications.

2.1 Assessment in primary school

Students are not graded in primary schools (grade 1 to 7). There are some assessments relating to mainly two purposes, early screening tests and national assessments.

Screening tests[7]: The major domains for early screening are the domains of reading and numeracy. They exist for grades 1, 2 and 3 – some are compulsory while others are optional (participation to be decided by the local authorities). In addition there is a screening test in English at grade 3 and one for ICT-skills for grade 4. With the ICT assessment as an exception, all these assessments are paper-based, and with the English test as the exception, they all relate to one of the five basic cross-curricular skills (and not to specific school subjects).

The tests are scaled through a calibration study in a representative sample. The focal aim of this calibration study is to identify the cut-score for the 20th percentile. The idea is that these screening tests should be used for early identification of struggling learners. To optimize this function, the tests are kept relatively easy and with a left-skewed distribution. In addition it is worth mentioning that the numeracy test is intentionally designed with timed tasks to ensure that students are using the desired procedure. For instance, with no time limits students may use repeated addition for solving multiplication items. By limiting the time, students not completing in time may be suspected to use such ineffective strategies.

Beyond the initial calibration studies data are not systematically collected for these assessments. The local school owners (the municipalities) are left to decide how to collect, process, and make use of the data. In general little is known about how these assessments are activated as a resource to inform instructional or other decisions in the schools. In particular, given that the main aim of these tests is to identify "students at risk", studies seeking to validate the predictive validity of the tests are needed.

National assessments: In the beginning of grade 5 there are compulsory national assessments in reading, numeracy and English. The English assessment relates to a limited selection of curricular aims in the reference school subject, while numeracy and reading relate to curricula across all subjects (see earlier description of the five so-called basic competencies). All the assessments were initially paper-based, but they are now electronically administered.

The students' scores are placed on a standardized scale with national average set to 50 points and one standard deviation set to 10 points on the scale. For the numeracy and English assessments the scales where fixed in 2014 and with subsequent tests referring to the scale established this year. This is achieved through the use of anchor items (15-20 items). Similar methods are established also with the reading assessments, with year 2016 as the fixed point. Even though the scales for the national assessments are

[7] In formal documents in English the tests are labelled as «mapping tests».

continuous, the results are also reported to teachers and students by a coarser measure placing the student in one of three predetermined levels of performance, with cut-offs at the 25th and 75th percentile of the distribution.

Although the system feeds detailed information about each student to the teacher, the main use of these assessments is for monitoring at the level of the individual school, at the municipality level and at the national level. The assessments can be regarded as low stakes for the students, but schools may perceive them as having moderate to high stakes, depending on how the results are used in the local dialogue and quality monitoring and development process. To our knowledge the results are not directly linked to rewards or penalties of any material nature, but still there are cases of local and national media presenting the results by identifying top- and low-ranked schools. Again, little or no research on psychometrical aspects of these assessments finds its way to scientific journals. The research that exists mainly relate to how stakeholders perceive and make use of the data[8], and some published studies have used data from one or more of the assessment to address substantial research questions not discussing or producing knowledge about the assessments as such[9].

Assessments supporting learning: While the most prominent and widely recognized assessments in primary school are the two types of tests described above, there is also a range of nationally developed assessments for various domains to be used by the teachers as tools in a formative context. Participation is voluntary (for teachers/schools). As for the screening tests, the final versions of the test forms are standardized in a nationally representative sample with the purpose to assist local interpretations of the students' scores. Beyond this sample, data are not systematically or routinely collected, and results are only used locally. Several test forms are developed for each of the domains, and they are all published in a web-based portal and made available for use for several years.

The three types of tests mentioned above are all developed by professional test developers employed at various academic institutions. The test development is based on frameworks with varying degree of operational fidelity and theoretical rigour. The assessments are also carefully designed with piloting and psychometric evaluations, applying both CTT and IRT. However, the analyses are published in technical reports with few, if any, papers published in peer-reviewed scientific journals. More ambitious and rigorously designed validation studies have to our knowledge not been conducted for any of the assessments.

Locally administered assessments: In addition some municipalities, Oslo in particular, have developed additional tests for local monitoring purposes. There are also a range of tests offered by textbook publishers and other private organisations. In particular many of these tests provide data for screening purposes in primary schools. Arnesen, Braeken, Ogden, og Melby-Lervåg (2018) conducted a systematic review of the documentation available for the tests most frequently used for screening of students social

[8] See for instance Mausethagen (2013) and Seland & Hovdhaugen (2017)
[9] See for instance Iversen & Bonesrønning (2015) and Roe & Vagle (2012)

functioning and reading proficiency and their general conclusion was depressing: The documentation is either completely lacking or very limited for most of these tests

Student survey: A national student survey is taken at the end of primary school in year 7 by all students. They provide data on their learning motivation and well-being as well as on classroom climate and the quality of their learning environment including teaching practices. Together with the various types of assessment data, this opens up for a broad range of applied research questions. Students responses to this survey is anonymous and it is not possible to merge these data to other student level variables. However aggregated data at the school level are available.

2.2 Assessment in secondary school

National assessments in lower secondary school: As for primary school, there are national assessments for grade 8 (and with the same test form administered also for grade 9 students the same year) in the same three domains as mentioned above. The same score reporting is used with a 50/10 scale. The only difference from the description given for grade 5 above is that these scores are reported to the students in the form of 5 separate levels (with cut-offs at the 10th, the 30th, the 70th and 90th percentiles). In addition, as for primary schools, assessments aimed at supporting learning, and some also for supporting grading, are nationally developed and administered through the national test portal, but they are not mandatory, and no data are systematically collected beyond the initial calibration study.

Marking in subjects: The most significant difference when advancing from primary to secondary school is marking. Students are graded by their teachers in all subjects in secondary schools. In lower secondary (years 8-10) there is a scale from 1-6, with no fail mark. In upper secondary school (years 11-13) a similar scale is used, but with the addition of the mark 0, and with 0 and 1 defined as fail marks. The mark for a subject in the final year[10] ends up in the report card. These report cards consist of teacher given end-of-year marks ("standpunktkarakter") and marks on a selection of exams. The marks are used for selection to the next educational level, that is average marks from lower secondary are used for selection to upper secondary, and marks from upper secondary are used for selection into higher education.

Teacher developed tests dominate: The prevailing testing and assessment practice at the secondary level are tests developed by the teachers' themselves. These tests are high-stakes in the sense that they are used for grading. To our knowledge no research or other knowledge production aiming at characterizing and understanding this practice exist. We do not know anything about the frequency, the format or nature, the inferential logic applied in marking or about the psychometric qualities of these tests. The variation along

[10] In lower secondary the vast majority of subjects are taught all years, and as such the grades received by the end of year 10 are the final grades ending up in the report card. Most likely, this also implies that students perceive the grades they receive in years 8 and 9 as having significantly lower stakes than the grades received in year 10. In upper secondary several subjects are finalized each of the three school years, with the result that the grades included in the final report card is a mixture of grades received over a longer time frame (but still with the grades received in the last year dominating the final report card).

such characteristics of this practice is in all likelihood very large. In addition to completely idiosyncratic practices, national level data suggest that there are systematic variations across subjects when it comes to grading. In a yearly statistics publication, "The Education Mirror", average marks in subjects are regularly reported – and some large differences are observed across subjects[11]. For instance, the teacher grading in English in lower secondary schools (end-of -year 10 mark) results in grades that are approximately half a grade above the average grade in mathematics. Differences across subjects, such as in this example, are not discussed or problematized to any large extent.

Long standing exam tradition: In addition, the Norwegian secondary school is based on centrally developed and administered exams. Although the specific practices of these exams have changed over the years, the function and general organization of the exams is part of a long-standing practice with historical traditions. Written, oral and practical exams are organized by the end of lower and upper secondary schools. Essentially the features of the exams are as follows:

- Lower secondary school:
  - Centrally developed written exams are administered for the subjects of Norwegian, Mathematics and English in lower secondary schools. Each student has to sit one exam in one of these subjects by the end of their 10th grade. In effect, one third of the student population are administered each of these tests every year.
  - In addition, each student has to sit one locally administered exam. These exams are administered for all subjects, and they are labelled as "oral exams", but may also involve written and/or practical/performance-based assessment components.
- Upper secondary school: In general there is a large variation in how exams are regulated across the various study programs. However, some general principles can be formulated
  - Centrally developed written exams are administered each year for all core subjects in the various programs (a large number of exams).
  - For most exams, a (random) selection of students participate.
  - Most exams are organised by the end of year 13, the final year in upper secondary school, but some exams are also administered by the end of year 11 and 12 for the subjects being finalised these years.
  - For the students in the general academic tracks:
    - Year 11: A small group of students (20 % of the cohort) are selected to take part in an oral-practical exam
    - Year 12: All students have to sit one randomly selected exam from the large selection of subjects. This may be a centrally developed written exam or a locally administered oral-practical exams – depending on the subject.

[11] See http://utdanningsspeilet.udir.no/2016/en/ for the 2016 report. All reports since 2004 are available online.

  - Year 13: All students have to take a written exam in the first language subject. In addition all students must have three additional exams, predominantly written and centrally developed exams.
- For the vocational tracks, students have to pass a performance-based and inter-disciplinary task simulating a real case from their profession (marked as either fail, passed or passed with distinction).

For all these exams there are a number of features varying over subjects with for instance preparation time, use of supporting materials and tools during the exams, and freedom to choose one or a few tasks from a list of suggested tasks. Little knowledge exists about how these features affect the quality of the exams. You may for instance suspect that the option to choose tasks and exam items from a list would make comparisons across students challenging and with the additional risk of introducing construct irrelevant variance. All written exams are marked by two external raters. In a meeting they reach agreement on the final mark given. In general, little is known about the rater agreement in this process. In general, it may also be noted that there are some large differences in average marks across the subjects, and these are in some cases contrary to what could be expected[12].

Student survey: Similar to the end of primary school, all students take the national student survey at the end of secondary school in year 10.

2.3 Standard-setting in the Norwegian context

Standard-setting is the concept that includes the definition of proficiency levels and corresponding cut-scores[13]. Pass-fail decisions are probably the best-known example. If 100 points are required to pass a test, 99.5 points mean that a student fails. Distinguishing not only between pass and fail but in addition between different passing levels such as sufficient, intermediate and advanced are sophisticated versions of standard-setting.

The main purpose of setting cut-scores is to define whether examinees reach specific proficiency levels or not and to communicate whether educational goals and expectations are being met or not. Thus, feedback to policy makers, schools and teachers is provided about strengths and weaknesses of a school system as well as about school and teaching quality including which individual students are at risk to fail[14].

[12] For instance, students choosing to study physics are on average among the highest performing students with overall high marks in the compulsory school subjects. However, the marks they receive in the subject of physics, both the teacher given end-of-school-year mark and the exam mark, are at par with or even lower than the grades awarded by other students with different specializations. In effect, since all these students, irrespective of choice of subject, can apply to many of the same higher education institution, students opting to specialize in physics are penalized relatively to students opting for other subjects.

[13] Cizek (2012)

[14] For more information see Blömeke & Gustafsson (2017)

The national curricula define only broad overall objectives for each subject and do not distinguish between proficiency levels. However, as should be realized from the above descriptions of the various assessments, standard-setting decisions are being made in several respects: for the screening tests a cut score for the weakest performing kids is defined, for the national assessments 3 and 5 performance levels are defined for year 5 and 8, respectively, and in grading decisions at the upper secondary school there is an operational pass/fail mark. No explicit (by authorities) or implicit (by teachers) link between the national curriculum, proficiency levels and grading is made. For some of the basic competencies standards are defined by describing levels of performance, but no empirical link is established between these largely normatively based statements about expectations and students' scores in a corresponding assessment. Since standard-setting is one of the most important and consequential decisions made in education, including normative-political, content-related and quantitative-methodological aspects, examining its reliability and validity would be crucial.

2.4 International assessments in the Norwegian context

Norway participates in a large number (in fact in most) of the international assessments. The table below summarizes the studies Norway has taken part in. Data for all the studies are freely available through the organizations' data repositories. In addition I may be mentioned that a majority of the studies have been conducted in Norway by a group at the University of Oslo having a close relationship with CEMO. This facilitates access to all types of materials from the studies.

Table 1. International large-scale assessments Norway has taken part in

Survey	**Domains/subject areas**	**Grade/age**	**When**	**Organization**
CIVED/ ICCS	CIVIC Education	Grades 8 (9)	1999, 2009 and 2016	IEA
ICILS	ICT literacy	Grade 9	2013	IEA
PIRLS	Reading	Grade 4 (5)	Every 5th year since 2001	IEA
TEDS-M	Mathematics teacher students competencies		2008	IEA
TIMSS	Mathematics and science	4 (5) and 8 (9)	Every 4th year since 1995 (Norway did not take part in 1999)	IEA
TIMSS Advanced	Mathematics and physics	12 (13)	1995, 2008 and 2015	IEA
IALS/ALL/ PIAAC	Adult literacies (reading, numeracy and ICT-related problem solving)	16-65 years	1998, 2003 and 2012	OECD

PISA	Reading, mathematics and science literacy	15-year-olds	Every 3rd year since 2000	OECD
TALIS	Teachers’and school leaders’ working environment	Lower secondary	Every fifth year since 2008	OECD

In general, a large number of studies using data from the international studies are published by staff at CEMO and other Norwegian researchers. Several PhD-theses have also been defended with a specific perspective on educational issues in Norway using data from these studies. A general conclusion is therefore that data from the international assessment to a large degree have been used for addressing a range of substantive research question, and data from these assessments are by far the most used resource for research by all the various assessments addressed in this document. However, there is a lack of research explicitly relating these assessments to other sources of information for students and schools in Norway (see also 2.5). The main reason for this disconnect is that person protection regulation prohibits data from these studies to be merged with other data for the participating students or schools. Hopefully, this can be made possible in future studies.

2.5 Register data

Norway and the other Nordic countries have very well-developed data registers with a huge collection of individual and school/institution level data. Statistics Norway (SSB) is in charge of building and maintaining these databases. Relatively simple aggregated and anonymous data are published as tables via a web-service at Statistics Norway[15]. In addition, The Directorate for Teaching and Training publishes school level data in their own data-base skoleporten.no.

The registers are not made available for the public. However, researchers can apply for access to anonymised individual level data. A detailed procedure for being granted access to such data is clearly described[16]. A specific database exists for educational data in the so-called National Education Database. Detailed descriptions of the content of the database are only available in Norwegian[17], but a shorter and more general statement is provided in English[18]. The database includes a wide range of data for each student/school. For instance students’ scores on the National Assessments and grades are available. Moreover, students can be tracked over time giving access to longitudinal data, and the multilevel structure with school and municipality is captured by the database. In addition, the data in the National Education Database can be merged with all other person level data in the SSB databases, for instance the students’ current educational or work life status, income, health or crime related variables etc. Unfortunately, item

[15] See http://www.ssb.no/en/
[16] See http://www.ssb.no/en/omssb/tjenester-og-verktoy/data-til-forskning
[17] See http://www.ssb.no/omssb/tjenester-og-verktoy/data-til-forskning/utdanning/nasjonal-utdanningsdatabase
[18] See http://www.ssb.no/en/omssb/tjenester-og-verktoy/data-til-forskning/utdanning

level data from the various assessments are not available. Furthermore, the instructional unit, the classroom, is not represented in the database.

Data from the registers may support a range of analyses of sources of variations at the individual, school and regional level, including individuals trajectories over time (but limited only to the marks reported on students final report cards in grades 10 and 13 as well as the three national assessments in numeracy, reading and English in grades 5, 8 and 9). For the national assessments access to individual and item-level data may be granted by the Directorate for Teaching and Training, but they are not allowed to store such data for more than a year or two in order to report back to schools, teachers and students.

CEMO is currently seeking to order a comprehensive and generic database form Statistics Norway. This database should support work with both master and PhD theses at the center. In order to get access to such a database ethical approval needs to be in place first, and as a general principle such approvals are provided for specifically formulated and delineated research projects. It is therefore still a bit uncertain whether it will be possible for CEMO to order a dataset to be used for such a generic and non-limited purpose.

3. Potential research areas for research on assessment in the Norwegian context

In the following we sketch up some potentially relevant areas for research on the national context of assessment in Norway. As stated in the introduction, the ambition is to further develop and establish projects with external funding based on one or two of these topics. However, the initial purpose is to provide information and guidance to potential applicants to the 2 positions as PhD at CEMO announced in the spring 2017.

In general, the interesting PhD research proposals will relate to educational measurement and quantitative analysis of a wide variety of data sets. As a PhD-candidate at CEMO you may well target substantial research questions in the field of education or questions targeting more generic methodological issues. However, some delimiting characteristics of the prospective research as part of a PhD-project at CEMO should be noted:

1) First of all, the research for these announced PhD positions has to include the Norwegian context. Knowledge of the Norwegian language is not a prerequisite because well-developed documentation of the data is to a large extent available in English, but for some studies mastery of Norwegian would be needed (in particular studies involving also the collection of additional data).
2) CEMO also supports research of a generic methodological nature without relation to the Norwegian context in other projects. This implies that if your intent is to address a generic

methodological topic, where the national context is not an integrated component of the proposal, your proposal would fit better to the other PhD-positions regularly announced at the centre.

3) No matter whether the research proposal is methodological in nature or relates to substantial research questions in education, it has to include the application of advanced quantitative/psychometric methods. The preferred applications, also in the case of proposals studying substantial research questions from the field of education, will be those which may be rated as methodological ambitious, and where some meta-perspective on methods, analysis and/or design is a part of the project. This does not exclude qualitative approaches, for example in terms of cognitive labs during test development, but in general we do not find a purely qualitative research proposal to be suitable for informing the field of educational measurement. In conclusion, the preferred applications for the positions will include assessment and/or measurement related issues as substantive parts of the intended research proposal

In the following some possible topics for research are briefly sketched out. For most of the topics a few possible research questions are also sketched out. The level of details in the research topics are currently not equal, which is a reflection of the fact that this is a dynamic documents. The list of topics and research questions are not meant to be exhaustive of the potential research questions to be asked in the announced PhD positions. They are meant as suggestions for potential candidates which may be elaborated in the brief (2-3 pages) document that they are required to deliver with their application. The defined areas are also to a large degree overlapping, so projects combining ideas across the nine numbered topics may well be formulated.

3.1 Linking assessments over grades and years to measure students' progress

In 2016 a white paper on the future of schooling in Norway[19] concluded (among other things) that curriculum and assessment need to develop descriptions and measures of students' progression. This is a concept that resonates well with vertical scaling/linking in the psychometric literature[20]. The limitation for research proposals relating to this topic is that relevant longitudinal data are not easily and readily available. The collection of such data may take more time than what is typically possible in the timeframe of a single PhD project.

However, following the principles of a so-called multi-sample sequence design, partial overlap in data collection over a few years may be used to model progress over a longer time frame. One example of such a study is the very large and ambitious German National Education Panel Study (NEPS)[21]. Currently, CEMO does not have the resources available for conducting a study of this scale, but by carefully selecting samples

[19] See http://nettsteder.regjeringen.no/fremtidensskole/files/2015/06/NOU201520150008000DDDPDFS.pdf for the full report in Norwegian.

[20] For a general introduction to the topic of linking and equating see for instance Kolen & Brennan (2014), and for a more specific discussion of vertical scaling of learning progressions or learning trajectories see Briggs & Peck (2015) and associated replies in the same special issue

[21] See https://www.neps-data.de/en-us/home.aspx

and designing new assessments with partial and systematic overlap in adjacent grades, it should in principle be possible to link data over longer periods of time. Building on the already existing national assessments for 5th, 8th and 9th grade, for example, adding an assessment in grade 7 with item sets overlapping with 5th grade on the one hand and 8th grade on the other hand would allow for covering progress in, for instance, reading or mathematics over a period of four years from 5th to 9th grade. Details of such a design needs to be further specified, and a close cooperation needs to be established with the institutions developing the assessments. Possible research questions are:

- RQ1: Is vertical scaling of the existing measures in reading, mathematics and English across 5th, 8th, 9th (and other grades) conceptually defensible and empirically feasible?
- RQ2: Is it possible to model students reading or numeracy skills (as operationalized by the national assessments) along one unidimensional scale, or are there infliction points in the development where the construct becomes more fuzzy or even separates into identifiable multiple dimensions or learning trajectories?

3.2 Modelling the change of institutions over time and how such changes affect achievement

The study of educational progress may also encapsulate research questions about how the school, the municipality or other institutional levels influence or relate to students achievement. This issue may be approached from several different viewpoints.

i) Multilevel and regression based analytical approaches may be applied to model how classroom, school- and municipality-level variables (and the interaction between variables at this level and variables at the level of the individual students) predict or account for variance in students' achievement. A concept relating to this approach is so-called value-added models[22]. The main purpose of value-added modelling is to compare each school (or other institutional levels) by adjusting or controlling for student characteristics (for instance variables describing students' home background). The idea is that these are given conditions beyond the control of the school, they are known to be strongly related to students' achievement, and they are typically unequally distributed across schools. But what if all schools had a comparable student mass? This is the question value-added models seek to answer. The aggregated achievement measure for each school provided by the value-added model would, under ideal conditions, be interpreted as a measure of school quality, thus enabling the identification of effective or "productive" schools. Available designs for modelling the value-added from schools range from simple linear regression with cross-sectional data to sophisticated multilevel and multigroup

[22] See for instance OECD (2008)

growth curve models where longitudinal data are used[23]. The analyst may go one step further to study how policy or school leadership malleable variables relates to the value-added.

Two national projects have published value added indicators at the school level – one for primary and lower secondary[24], and one for upper secondary schools[25]. The aim for both projects has been to study how the current register data can be used to estimate school quality, or more specifically, the contribution of the school to students learning. It is interesting to note that the models used in these two projects are not identical, among others with respect to the dependent variable. Both reports have stirred a (largely ideological) debate about the validity of using these (or similar) indicators as a measure of the quality of the educational provision in schools and municipalities, and also internationally, particularly in the US, there is heated public and academic debates in relation to so-called teacher added-value models used to evaluate teachers.

ii) Another related, but still different line of research is not mainly concerned with identifying effective schools, but rather to model how changes over time at the institutional or system level (schools, municipalities or countries) relate to or affect students' achievement in the same period. Potentially important omitted variables also change over the same time period, and, in order to decrease or wash out the effect of such exogenous variability, methods such as difference-in-difference, regression discontinuity, propensity score matching, instrumental variables, or other methods from econometrics, are possible analytical approaches to study these types of institutional level longitudinal data. The challenge of this line of research is that the data currently available in the national registers at the school level is largely limited to variables describing resources available at the school (number and composition of staff, availability of computers, indicators of school and or municipal economy etc.). Characteristics describing the processes ("what happens") in schools are currently not available. Some of these variables (e.g., instructional quality, teaching practices, classroom climate) are however collected in the international assessments and the student surveys implemented in years 7 and 10. It is therefore important to find ways to either merge data from the national registers or to collect new data containing also these types of variables over time for the institutions. The latter would of course pose challenges for a 3 to 4 year PhD-project.

Possible research questions for this line of research could be:

[23] In many ways, such value-added models are justified in the absence of vertically linked scales for student achievement (see 3.1). If vertically linked scales had been in place, the value added could simply be represented as the change in the score on the scale.

[24] See https://www.ssb.no/utdanning/artikler-og-publikasjoner/er-det-forskjeller-i-skolers-og-kommuners-bidrag-til-elevenes-laering-i-grunnskolen for the full report in Norwegian. A very brief English abstract is included in the report.

[25] See http://www.sof.ntnu.no/SOF_R_01_16.pdf for the full report in Norwegian

- RQ1: What is required for developing an efficient and unbiased estimate of school quality? What are the major methodological challenges in estimating value-added indicators for Norwegian schools and municipalities?
- RQ2: Does the current available database of register data for schools and municipalities reflect current theories and previous research on the characteristics of successful schools? Given the data in the registers, which characteristics account for the variability across schools and/or municipalities?
- RQ3: Given the successful development of a school leader questionnaire, and subsequent successful data collection, how can approaches as those mentioned above inform us about possible school developmental trajectories?

3.3 Validity of low-stakes assessments for system-wide monitoring

International and national assessments are actively used for monitoring purposes in Norway and many other countries. Interpretations of such measures usually will assume that they reflect students' true abilities. However, one threat to the validity of such claims is the fact that these assessments are low-stakes assessments with no immediate or direct consequences for the students, and in the case of international assessment, results are typically not fed back to either the schools or students. Moreover, the relationship to curriculum is probably perceived to be somewhat weak: In the case of the international assessments blueprints are developed through some consensus process among the participating countries leading to assessments with partial correspondence with the actually taught curriculum, and in the case of the national assessments the measure covers only a small part of one subject (English) and the two others (reading and numeracy) measure cross-curricular generic competencies. These phenomena may lead to variability in students' engagement with the assessments, and the level of engagement may vary systematically across groups of students or across schools, leading to what is known as systematic measurement errors (construct underrepresentation and/or construct irrelevant variance). These and a range of other threats to validity of the interpretations currently being made from these assessments need systematic study. On this background, the following are some of the many research questions which may be posed:

- RQ1: Which educational decisions are based on these measures (at student, classroom, school or municipality level), and how valid are results from low-stakes assessments for informing such decision making?
- RQ2: Are students equally motivated to perform at their best in the test situation across schools, socioeconomic strata, gender or countries?
- RQ3: Do the many types of assessments implemented in Norway have sufficient psychometric quality to support some of the main uses of the scores? (This perspective includes many different questions such as measurement invariance across different groups (gender, language background etc.), whether items show differential item functioning or how the linking across cohorts works out).

- RQ4: Particularly for the screening tests: Do they effectively identify students at risk. That is: Do they predict outcomes also at much later points in time, for example as shown by the national assessments, exams or school marks, or even longer term outcomes such as participation in the work force, salary, health or other outcomes in life?
- RQ5: In cases involving standard-setting, what are the psychometric qualities of the cut-scores? Would different methods for developing the cut-scores result in the same cut-scores? Would different groups of experts come to the same conclusion when assigning items to proficiency levels (teachers, policy makers, students, researchers)? What type of students do they have in mind when they are classifying (top, medium, weak students, students applying different learning strategies)?

3.4 Using register data for educational research in Norway

As briefly sketched out above, Norway has a well-developed system with register data where the individual students and schools can be linked and monitored over time. In general this topic relates very much to most of the other topics identified in this list. In general we would welcome projects formulating all kinds of substantive research questions utilizing the available register data by applying advanced quantitative methods, but also research with a more meta-perspective on the potential of using register-data for a range of purposes.

- RQ1: How does the Norwegian national registers compare to the other systems in use? Do they provide adequate support for targeted and relevant analyses of policy malleable features? What are recommended improvements? This is a more evaluative question, but relevant studies would include systematic reviews of historical and current use of the National Education Database, and comparisons with other countries' registers.
- RQ2: Taking a longer historical perspective: How has Norwegian schools/students developed over the last 20-30 years (not only in terms of averages and other simple descriptive statistics, but also in terms of larger structural relationships with hierarchical data-structure etc.)? The implementations of reforms are examples of critical time points where important system level characteristics may gradually be changed.
- RQ3: How can register level data be used as a source for developing improved longitudinal profiles of students' competencies or schools developmental trajectories over longer time frames?
- RQ4: What is the potential in merging existing register data from education with other registers such as those describing early years, health, family, work, etc. (potentially in cooperation with another existing large-scale project, the SOL-study[26])?

3.5 Linking the national system with the international studies

In Norway, as in many other countries, the results from the international assessments are used, or at least referred to, when policy changes are suggested and finally decided and implemented. The literature on the opportunities and limitations of basing policy making from practices in other countries is abundant, ranging

[26] See https://www.fhi.no/en/studies/language-and-learning-study/

from positive views about how Western schools may learn from practices in the Eastern Asian top-scoring countries, to similarly sceptical literature warning against and discussing why such policy borrowing is difficult or even futile.

In order to improve the valid interpretation of data from the international assessments, it has been suggested that countries would need a strategy or a framework for validation[27]. Carefully designed studies are needed in order to address those aspects of the studies which are seen as particularly important or influential for national evaluation and policy development. The challenge for several of the possible questions posed below is that data from the international assessments are anonymous. Potential relevant research questions for this topic are for instance:

- RQ1: How is the fit between what the international studies measure, and what students are supposed to learn in school? How can measures of alignment or opportunity-to-learn influence interpretations of results from the international studies[28]?
- RQ2: How well does student self-reported data capture the register data available for the same phenomena – for instance students self-reported indicators of socioeconomic background (SES)?
- RQ3: Using available register data for the same cohorts as those participating in one or several of the international studies, how comparable are the results, for instance with respect to the relative distribution of variance to schools, gender, SES, immigrants, etc.?

3.6 Studies of the national system of exams

As indicated in the general description of the Norwegian school system above, we lack knowledge about psychometric qualities of the most important assessments in Norway: Teachers' own tests and the national exams. The next topic (3.7) relates to the teachers' assessment practices, while the following motivates and describes research relating to the exams.

The scale of the procedures relating to the marking of exams is already very large – they are after all tests that are developed and administered every year for large samples or even complete cohorts in a large number of subjects. The administration of the yearly exams is probably one of the most resource intensive tasks carried through by the educational administration from national to local level in the Norwegian school system, and huge resources are invested in quality ensuring the whole process. For instance all the hundreds or thousands of exams administered every year are rated by two external and independent raters. However, the observations (and potential data) from this large-scale procedure are not collected in any systematic fashion. The only data being recorded is actually the final mark for each of the students – one number for each student for each exam. In other words, studies of the psychometric quality of exams are only possible if relevant data are collected. CEMO will therefore take initiative for collaboration with the National Directorate for Teaching and Training in designing a data collection procedure that would

[27] Rutkowski & Delandshere (2016)
[28] One case can be found in Pedersen (2013)

support a range of research studies on the quality of the exams and which at the same time would be perceived as supporting raters' work without introducing a new large burden for the raters, and without introducing new huge costs. Some procedures like this are already in place for some exams in the science subjects in upper secondary schools.

In general, we are optimistic that designing such a data collection should be possible. Some data can be collected simply by storing them, namely the two raters' initial marking of the exams. It might be that the design of how students are crossed within raters is not optimal in the existing system, but this could also easily be changed without introducing the development of new tools and without introducing increased costs. However, to be able to conduct more detailed analysis, more detailed rubrics with the use of analytic scoring needs to be developed, at least for school subjects where the existing exams represents test forms built by a collection of a larger number of independent items. In these cases the raters are possibly already using some form of analytical scoring rating each item independently. In addition some person level data need to be stored allowing for merging with register data, such as gender and indicators of socioeconomic background. It is sufficient that these data are collected only for a modest sample of students.

Both, the lower- and the upper-secondary examination systems are sample based in the sense that only a randomly selected share of students takes part in each subject. Since average marks from lower secondary are used for selection to upper secondary, and marks from upper secondary are used for selection into higher education, fairness and validity questions of such a system are crucial. Furthermore, students have limited time to prepare for the exams they were selected for and they can bring a broad range of materials with them. Concerns exist that these characteristics may favour students from academic families (Nusche, Earl, Maxwell, & Shewbridge, 2011).

A few potential research questions are given in the following:

- RQ1: How does the introduction of joint marking guides or rubrics at the level of the single item affect how exams are rated? Are the raters willing to comply with the rubrics? Does the introduction of rubrics increase reliability of the exams and/or does it affect the mean scores and variability of the scores (within and across schools)?
- RQ2: How well does existing exams live up to ordinary psychometric requirements for tests (DIF, dimensionality, model fit etc.)?
- RQ3: How does the scores on exams relate to teachers' marks for the same students in the same subjects? How comparable are the standards used for exams across subjects (i.e. are exams for some subjects relatively "easier" given all other information we have about the students)?
- RQ4: Which inferences are supported by the exams? Do they support intake to higher education across cohorts? Do they have predictive validity for specific studies beyond what can be obtained from the grade point average? Can the results for one cohort be compared with the results for other cohorts?

- RQ5: Given the random assignment to exam subjects and types: Do all students have the same chance to show their best performance? Does the relation between exam grades and family background differ compared to teachers' marks or other assessment outcomes?

3.7 Teachers' and school leaders' assessment literacy

The contested and much discussed introduction of system-wide assessment for accountability and/or monitoring purposes is not only associated with the idea that test scores provide incentives for schools and students, but also the more formative purpose of informing instruction in classroom and practices at school and system-level by data[29]. One premise to make such assessment-driven systems work is that all the actors are able to make use the data for the range of purposes they are intended to be used.

According to a framework proposed by Gummer and Mandinach (2016), such assessment literacy consists of four types of skills: (a) to understand the outcomes coming from external assessments such as national or international studies; (b) to create classroom assessments on the basis of basic psychometric principles; (c) to interpret the outcomes of one's own designed assessments; and (d) to draw conclusions on instruction and the improvement of schooling in general.

Studies of how well teachers are prepared to develop, mark, score and make use of their own assessments and how competent they are in interpreting students' scores on external tests is a growing and relevant field of research. Studies in this field also include measuring perspectives, preferences or attitudes relating to assessment that teachers brings into their work. Critical reviews of the coverage and psychometric quality of existing instruments highlights the need for continued work with validating the existing and or developing revised or new instruments[30]. In the Norwegian context, conducting a study in this field would also for the first time provide information about teachers' assessment literacy. A closely related field of research which is also growing are studies on so-called data-based or data-informed decision making. This also includes the study of how well prepared teachers, school leaders and other stakeholders are in analysing data and using the types of data that are regularly and increasingly reported to the schools based on national assessments and similar large-scale assessments used for monitoring and supporting local work with school improvement. Some tests are available here as well that seek to measure a broad data-use concept[31] or more narrowly defined constructs which could be labelled as statistical[32] and research literacy[33].

Some highly relevant qualitative research has already been conducted in Norway, mostly through studies where teachers and other stake-holders in education are interviewed and/or observed in some select

[29] For a discussion of some of the issues involved see http://www.rand.org/content/dam/rand/pubs/occasional_papers/2006/RAND_OP170.pdf
[30] Gotch & French (2014); DeLuca, LaPointe-McEwan, & Luhanga (2016)
[31] Jimerson (2016)
[32] Stone (2006)
[33] Groß Ophoff, Schladitz, Leuders, Leuders, & Wirtz (2015)

situations[34]. In general there is also a large amount of studies relating to teachers' formative assessment practices. This is also to a large extent research based on qualitative data. However, no study involving the measurement of teachers' or teacher students' actual assessment literacy has, to our knowledge, been conducted in Norway.

Moreover, the actual practices of teachers when it comes to assessments with a summative purpose are in general not much studied – not in Norway and not to a major extent internationally. Particularly in a system like the Norwegian, where teachers are trusted to grade their own students without any external inspection, monitoring or moderation, knowledge about this assessment practice is indeed needed. There are no existing data which may be used for this line of research, and the prospective PhD project within this topic would also need to develop a larger data collection design. Relevant data to be collected could be teachers' records throughout the year with marks on specific tests or other performances/observations. In addition the actual tests or other instruments should be collected and coded. This would help build up a database of teachers' own assessment practices to be used for a range of analyses aiming at giving a representative view of teachers' summative assessment practices in a few selected school subjects.

Finally, as mentioned grading starts only in year 8. Still, teachers have to diagnose and to evaluate student achievement and to communicate strengths and weaknesses to the students themselves, to their parents, to colleagues and other parties involved. Information about the assessment practices, the criteria used (individual progress, norm- or criterion-referenced) in this context is – to our knowledge – missing. Furthermore, it is an open question how students and parents perceive the evaluation they receive, whether they are able to get an idea of where the student stands.

- RQ1: What characterizes teachers' assessment practices in terms of tests for grading purposes (summative evaluations)?
- RQ2: What are important characteristics of their assessment instruments (item types, formats, test formats and modes, length of tests, feedback format, etc.)? Are concerns about reliability and validity raised in the process of developing assessments?
- RQ3: What characterises teachers' views on grading and assessment, and what characterises their assessment literacy measured as a competency?
- RQ4: How do teachers and schools actually use the information provided by the many different external assessments? Do various stake-holders have what could be called data-literacy?
- RQ5: How reliably and validly do teachers summarize their evaluations of student achievement in primary school where grading does not take place? Do students and parents recognize strengths and weaknesses appropriately?

[34] One such study is the PraDa study. This link is to a page in Norwegian, but there is a list of publications of which some are in English: http://www.hioa.no/Forskning-og-utvikling/Hva-forsker-HiOA-paa/FoU-SPS/prosjekter/Bruk-av-elevresultater-data-i-norske-skoler-og-kommuner-PraDa

References

Arnesen, Anne, Braeken, Johan, Ogden, Terje, & Melby-Lervåg, Monica. (2018). Assessing Children's Social Functioning and Reading Proficiency: A Systematic Review of the Quality of Educational Assessment Instruments Used in Norwegian Elementary Schools. *Scandinavian Journal of Educational Research*, 1-26. doi: 10.1080/00313831.2017.1420685

Blömeke, Sigrid, & Gustafsson, Jan-Eric (Red.). (2017). *Standard Setting in Education. The Nordic Countries in an International Perspective*. Rotterdam: Springer.

Briggs, Derek C., & Peck, Frederick A. (2015). Using Learning Progressions to Design Vertical Scales that Support Coherent Inferences about Student Growth. *Measurement: Interdisciplinary Research and Perspectives, 13*(2), 75-99. doi: 10.1080/15366367.2015.1042814

Cizek, Gregory J. (Red.). (2012). *Setting Performance Standards Foundations, Methods, and Innovations*. New York/London: Routledge.

DeLuca, Christopher, LaPointe-McEwan, Danielle, & Luhanga, Ulemu. (2016). Teacher assessment literacy: a review of international standards and measures. *Educational Assessment, Evaluation and Accountability, 28*(3), 251-272. doi: 10.1007/s11092-015-9233-6

Gotch, Chad M., & French, Brian F. (2014). A Systematic Review of Assessment Literacy Measures. *Educational Measurement: Issues and Practice, 33*(2), 14-18. doi: 10.1111/emip.12030

Groß Ophoff, Jana, Schladitz, Sandra, Leuders, Juliane, Leuders, Timo, & Wirtz, Markus A. (2015). Assessing the Development of Educational Research Literacy: The Effect of Courses on Research Methods in Studies of Educational Science. *Peabody Journal of Education, 90*(4), 560-573. doi: 10.1080/0161956X.2015.1068085

Iversen, Jon Marius Vaag, & Bonesrønning, Hans. (2015). Conditional gender peer effects? *Journal of Behavioral and Experimental Economics, 55*, 19-28. doi: http://dx.doi.org/10.1016/j.socec.2015.01.003

Jimerson, Jo Beth. (2016). How are we approaching data-informed practice? Development of the Survey of Data Use and Professional Learning. *Educational Assessment, Evaluation and Accountability, 28*(1), 61-87. doi: 10.1007/s11092-015-9222-9

Kolen, Michael J, & Brennan, Robert L. (2014). *Test Equating, Scaling and Linking*. New York: Springer.

Mausethagen, Sølvi. (2013). Talking about the test. Boundary work in primary school teachers' interactions around national testing of student performance. *Teaching and Teacher Education, 36*, 132-142. doi: http://dx.doi.org/10.1016/j.tate.2013.08.003

Nusche, Deborah, Earl, Lorna, Maxwell, William, & Shewbridge, Claire. (2011). *OECD Reviews of Evaluation and Assessment in Education: NORWAY*. Paris: OECD.

OECD. (2008). *Measuring improvements in learning outcomes. Best practices to assess the value-added of schools*. Paris: OECD Publishing.

Pedersen, Ida Friestad. (2013). Is TIMSS Advanced an appropriate instrument for evaluating mathematical performance at the advanced level of Norwegian upper secondary school? An analysis of curriculum documents and assessment items. *Acta Didactica Norge - tidsskrift for fagdidaktisk forsknings- og utviklingsarbeid i Norge, 7*(1), 24.

Roe, Astrid, & Vagle, Wenche. (2012). Kjønnsforskjeller i lesing – et dybdedykk i resultatene fra nasjonale prøver på åttende trinn fra 2007 til 2011. *Norsk pedagogisk tidsskrift, 96*(06), 425-441.

Rutkowski, David, & Delandshere, Ginette. (2016). Causal inferences with large scale assessment data: using a validity framework. *Large-scale Assessments in Education, 4*(1), 6. doi: 10.1186/s40536-016-0019-1

Seland, Idunn, & Hovdhaugen, Elisabeth. (2017). National Tests in Norway: An Undeclared Standard in Education? Practical and Political Implications of Norm-Referenced Standards. I Sigrid Blömeke & Jan-Eric Gustafsson (Red.), *Standard Setting in Education: The Nordic Countries in an International Perspective* (s. 161-179). Cham: Springer International Publishing.

Stone, Andrea. (2006). *A Psychometric Analysis of the Statistics Concept Inventory.* (PhD), University of Oklahoma.

Tveit, Sverre. (2014). Educational assessment in Norway. *Assessment in Education: Principles, Policy & Practice, 21*(2), 221-237. doi: 10.1080/0969594X.2013.830079