Policy Legitimation of Educational Assessment Reforms

The Cases of Norway and Sweden

Sverre Tveit

Thesis submitted for the degree of Philosophiae Doctor (PhD)

Department of Education, Faculty of Educational Sciences
University of Oslo

May 31st, 2019

Series of dissertations submitted to the
Faculty of Educational Sciences, University of Oslo
No. 311

ISSN 1501-8962

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Summary

This thesis researches how policymakers engage in international research and policy discourses in the legitimation of nation states' educational assessment and testing policies, and examines the associated tensions related to the purposes of assessment. The study analyses policy documents and policymaker interviews from Norway and Sweden as well as policy information from European and global research organisations and policy agencies. Expert interviews with policymakers at the political and administrative levels of the education ministries and at the associated executive agencies help substantiate how the governments give emphasis to various purposes of educational assessment. Given their political and high-stakes character, educational assessment and testing policies are suitable areas for investigating strategies and trends of policy legitimation in education. The thesis investigates the national testing programmes in both countries, the curriculum reform and associated grading regulations in Norway, and the reform of formal grading age policies in Sweden.

While the focus of the study is on contemporary policymaking, with an emphasis on the reforms of the past two decades, this is interpreted in the light of historical developments within and beyond the national contexts. As such, emphasis is placed on the national and transnational history of educational assessment, the role of global research and policy collaboration, and national policymakers' use of supranational policy agencies and research networks to legitimise reforms. While acknowledging that there are other spaces, actors and factors that shape governments' agendas, the thesis focuses on policymaking in the intersection of the national and the global in the legitimation of assessment reforms.

The thesis promulgates an analytical framework that distinguishes between educational assessment used to *certify*, *govern* and *support* learning and instruction, and discusses the emphases on these in the legitimation of nation states' policies. The thesis connects these three contemporary assessment roles to historical developments, identifying three transnational trends. While *meritocracy* was a key focus in the international research and policy discourses of the 1930s, the focus shifted to a greater emphasis on *accountability* policies from the 1990s, while *Assessment for Learning* became important at the turn of the millennium. The thesis highlights the accumulation of multiple roles of educational assessment in response to the transnational trend at the time of implementation, revision and expansion. Influenced by the meritocracy, accountability and Assessment *for* Learning trends, contemporary national assessment instruments have accumulated the roles of certifying, governing and supporting learning and instruction.

The thesis illuminates that while the national tests in Norway were initially designed to primarily serve a governing role, the subsidiary role shifted from certifying to supporting learning and instruction following the change of government in 2005. The emphasis on the national tests' potential use in formative assessment undermined the use of the tests as a governing instrument. The Swedish national tests, by contrast, combine all three roles of educational assessment, which has caused conflicts with respect to which role to prioritise.

The thesis furthermore identifies *three modes of policy legitimation* that shed light on how policymakers engage in policy and research discourses to legitimise reforms. The thesis portrays a shift away from *collaboracy*—defined as policy legitimation located in partnerships with and networks of stakeholders, researchers and other experts—towards a greater use of supranational agencies (described as *agency*) such as the OECD, the EU and associated networks, as well as the use of individual consultants and private enterprises (described as *consultancy*) to legitimate reforms. The thesis reveals a remarkable case of policy legitimation in Sweden, where the government nominated a neuroscience professor who consulted for the government on the proposed policy changes by referring to OECD reports and EU policy descriptions of European countries' formal grading age policies. By constructing "world situations" with respect to formal grading age, the government (mis)used comparative data to effectuate an assessment reform that sought to break with the Scandinavian tradition of prohibiting formal grading in primary education.

The thesis highlights the OECD's role in nation states' policymaking and portrays a Scandinavian *governance turn* that embeds several nuances. The thesis discusses how nation states' concepts of knowledge and skills reflect reciprocal processes of defining and determining goal and standard attainment between the nation states and transnational and supranational spaces. The thesis also points to how governments use formative assessment and Assessment *for* Learning policy and research discourses in reform agendas. Based on the reform of the assessment regulations in Norway, the thesis exposes definition problems and discusses problems related to borrowing and lending in the formative assessment research and policy discourses. It addresses the (low) comparability of the highly accountability and testing–oriented British policy context, and the Scandinavian policy context with less emphasis on accountability, resistance to standardised testing and a legacy of prohibiting formal grading. The thesis illuminates how the concept of formative assessment becomes diluted when the borrowing and lending of such a vague concept occurs across widely different policy contexts, and highlights that it is this vagueness that makes the formative assessment and Assessment *for* Learning policy and research discourses so powerful.

Acknowledgements

While taking the train from Kristiansand to Oslo to submit my PhD thesis, I carefully considered the many persons and institutions that have contributed to my over eight-year journey. First, I wish to thank my two supervisors, Berit Karseth and Christian Lundahl, who have continued to challenge and support me throughout this journey, which felt at times as though it would never reach its destination. I have finally arrived!

Christian has had the responsibility as the main supervisor throughout most of the project. Berit started as the main supervisor. The loss of Erling Lars Dale, the lead scholar on my topic in Norway, only half a year into the project was unfortunate. It was around that time I was introduced to Christian. It became clear that, given his expertise in the field of educational assessment in Sweden and beyond, I would benefit from having him as the main supervisor. When Berit was elected Dean for the Faculty of Educational Science in 2013, Christian was assigned the responsibility as the main supervisor.

Berit, it took me several years to fully embrace the institutional theory perspectives you encouraged me to adopt, and I was never an easy student. Thank you for your patience, for forgiving the many detours I followed and for your thoughtful feedback on my manuscripts over these years. I know of no better manuscript commentator than you, and I hope to have the chance of writing together someday! I am especially grateful for the time you spent supervising me this last year, so that I finally managed to complete this much overdue thesis.

Christian, I did not know you when I embarked on this journey, but I soon learned how fortunate I had been to find an expert east of the border who shares my passion for assessment policy. You have indeed inspired me to immerse myself in the historical roots of contemporary policies. Furthermore, you invited me to take part in two of your research projects: As a research assistant for a report to the Swedish Research Council entitled *The Geography of Grading – research on grades and summative assessments in Sweden and internationally*, and as a co-author of a report to the Swedish Agency for Education on *Grading systems in an international perspective*.

The data set of the report to the Swedish Research Council was the basis for the article I co-authored with Christian. This project inspired my development of the distinctions between *three modes of policy legitimation*, which became a backbone of my thesis. I am very grateful for your faith in me despite the many side-tracks and setbacks throughout these years. I truly feel that I have learned from the master of this field in Scandinavia, and I hope to have the opportunity to work with you to explore and reveal new territories of educational assessment in the future.

While I pondered adopting a mixed-method design we discussed the prospects of Rolf Vegar Olsen as a co-supervisor. However, limiting the scope of my thesis to expert interviews and document analyses was one crucial step in narrowing the scope of my study. I decided to collaborate with experts in other methods, rather than try to do everything myself. Therefore, it was with great enthusiasm that I accepted Rolf's invitation to co-edit a special issue of *Acta Didactica*, which even

led to a co-authored article based on my PhD data. The chance to collaborate with you, Rolf, reflects and reaffirms my strong commitment to bridging the education policy and educational measurement camps. And equally important, I look forward to peer review more local beers with you at the AEA-Europe conferences in the years to come!

I am also grateful to the Department of Education, University of Oslo, for providing the research facilities and especially to the Curriculum Studies, Leadership and Educational Governance (CLEG) research group for providing a scholarly community that supported my progress. A special thanks goes to Jorunn Møller and Kirsten Sivesind for challenging my thinking and for introducing me to the world of scholars through conferences and seminars. I also gratefully acknowledge the support of the National Graduate School of Education (NATED) in preparing me for the challenges of my research career and for providing financial support for my research stays in both California and Sweden.

Thanks also to Guri Skedsmso and Gunn Elisabeth Søreide for their comprehensive, supportive feedback on my manuscripts at the 50% and 90% seminars. I am also indebted to Uppsala University, especially Eva Forsberg, Wieland Wermke and other members of the Studies in Educational Policy and Educational Philosophy (STEP) research group for hosting me during my research stay in 2014–2015.

My deepest appreciation goes to the Department of Education at the University of Agder, where I have had a full-time teaching position since my PhD fellowship in Oslo ended in 2016. I am especially grateful to the head of my department, Inger Marie Dalehefte, for allowing me to focus on my PhD thesis this spring. I am likewise grateful to the members of our General Didactics and Education Policy research group (commonly known as Didaktikkgruppa), chaired by Turid Skarre Aasebø, for contributing to a stimulating intellectual environment and for many great parties! Ilmi, thank you for our many intellectual conversations at work, and in the bars as well! My special thanks go to Andreas Reier Jensen for helping me navigate rough seas, both at work and in life generally, after I moved to the south. I look forward to pursuing our planned research collaboration!

A warm thanks to my students in the Bachelor of Education programme at the University of Agder, especially the student representatives I have worked with as the Academic Programme Director the past two years. Our joint struggles to discuss the content, especially the learning theories in PED128, were instrumental for my development as a university teacher. I encourage you to continue your efforts and never stop challenging your teachers and your university. Union work was the best education I had, so promise me you'll get that union for special needs education and education students (SOPIA) going again!

I am thankful to Mark Wilson of the University of California Berkeley and to Thomas Hatch of Columbia University's Teachers College for inviting and hosting me for research stays in 2012–2013 and 2016, respectively. I am also indebted to the Fulbright Foundation for a research grant supporting my stay in California, to the Tokyo Foundation for the Ryoichi Sasakawa Young Leaders Fellowship Fund (Sylff) scholarship and Sylff Research Abroad scholarship and to Robert

and Ella Wenzin's endowment at the University of Oslo for financial support of my research stays in New York and Sweden.

I want to express my great appreciation to the Norwegian Directorate for Education and Training (DET), the Swedish National Agency for Education and the education ministries in both countries for allowing me to undertake interviews for this study. I especially value the outstanding collaboration with DET in regard to ongoing reforms of the examination system. I hope that policymakers and us scholars will continue to balance between criticism and constructive input that help improving policy and practice. I hope the thesis will be well received, both the criticism I put forward, but also the strengths I have identified in our assessment culture and approaches to policymaking. This thesis indeed highlight advantages of close collaboration between policymakers, researchers and the teacher profession – I even coined a term for it!

My parents deserve special thanks for patiently supporting me all these years. Mom, thank you for the many sumptuous dinners and for your dedication to education, which has obviously affected my career choice. Finally, I can wear the *bunad* next May 17th!

Dad, it was not easy to take over the family property and get more involved in the family firm at Gautefallheia in Telemark, in the middle of my toughest years in academia, but together we managed! It is unlikely that I had gotten this far in my commitment to improve education if it was not for your support after I dropped out of high school in 2000. Your support and flexibility made it possible to take the opportunities that came my way during difficult teenager years. This way you earned my loyalty and commitment to take *my share* of responsibility for the family business, during university studies and as a PhD research fellow.

Given the challenges our family have faced at times, I think we can be proud that we managed to collaborate in expanding the property development while at the same time I ultimately was able to complete my PhD. This accomplishment would not have been possible without the good collaboration with and loving support from mom as well.

A warm thanks to Katrin, our head of office for many years, for keeping up with my father and me throughout the firm's many ups and downs. I truly value how you understood and respected my priorities, Katrin, although they many times were at odds with the tasks you had to deal with in my absence. Likewise, I want to thank Roald, for respecting my priorities although you had hoped to see me more involved in the firm. Thanks also for many stimulating academic discussions!

A warm thank you is due the friends in academia who have accompanied me on this journey. Christina, Maike, Jeff, Sølvi and Tine, we were colleagues in the CLEG research group and in NATED, but we developed friendships that have continued to stimulate our intellectual work. Although we meet less frequently these days, I value these relationships and hope they will endure throughout our careers. I am especially grateful to Christina, for being such a good advisor about though priorities both at work and in life. Likewise, to Tine, for the close connection we have developed since our simultaneous research stays in Berkeley in 2012. I feel extraordinarily privileged to learn from your expertise and experience and I am grateful for the opportunity to

expand on our joint research interests through the inter-institutional research group Studies of the Teaching Profession, Teacher Education and Education Policy (TEPEE) .

My special thanks go to Judit Novak, whom I met in Uppsala and who has since been my closest friend in academia. Your intellectual capacity and scholarly passion have fascinated me since day 1, and I am honoured that you continue to share with me your talent and commitment to research. I hope that we can continue to inspire and support each other throughout our careers—even, finally, in the same country!

Cheers to my 'Australia friends' Helga, Jen, Kate & Andy, Leigh and my 'favourite English teacher' Rob; to my many 'Tübingen friends', especially Barbara, Vid, Jaroslav, Chris, Jo, Ben & Claire; to my 'California friends' David, Elizabeth, Johannes, Doruk, and especially Sarah, 'the Dane', with whom I share warm memories from 1503 MLK; to my 'Uppsala friends' Nils, Håkan, Andrés, Mette and Josef. Each and every one of you share a small piece in this accomplishment, as you – knowingly or not – assisted me in grasping the social, cultural, political and professional finesses of your particular country and city. Our conversations, trips, squash games and parties were the most formative parts of this journey - and were an important motivation when I immersed myself into the complex literature on comparative methods and comparative education.

I also wish to thank the friends I made 20 years ago during my time in the School Student Union of Norway (Elevorganisasjonen), who still remain my closest friends. I especially thank Marianne, Magnus, Christopher, Knut, Halvard and Robert for celebrating my achievements, helping me to get through my toughest days during these years and for many great holidays, evenings and cabin trips together.

Especially thanks to Halvard, for our many long talks about politics (in Oslo, nationally and internationally). As President for *Elevorganisasjonen*, you asked me to collaborate with the union in writing a report with the aim of making assessment policies and practices a political priority. Little did we know then, that it would accumulate to a 20-chapter book, published in 2007, with contributions from key scholars as well as two contributions from the union and three by myself. It was a pleasure to work with you and Ragnhild, and later Jan Christian, on that project. We succeeded only partly. Our call for better feedback practices in schools quickly gained momentum, manifested in comprehensive national Assessment *for* Learning policies and professional development programmes. With respect to the examination system and grading it took more than a decade longer. But now we are getting there!

Greetings to Terje, Linda and all the 'kids' at *Terje's skischule*, the annual skiing trip to the Alps; this winter I can finally honour the promise to leave my laptop at home! A big thanks to Magnus, Knut, Espen, Robert and Marianne for hosting me during my many visits to Oslo, Bergen, Svalbard, New York and Washington, DC, respectively.

The DC trips were also the starting point for a book Marianne and I will publish next year. Who would have thought, Marianne, that day we first met in March 2001, having just been elected to the central board of *Elevorganisasjonen*, that twenty years later we would be such close friends and even co-authors of a book on grading in schools!

Finally, I wish to acknowledge two outstanding persons who are no longer with us.

**

Erling Lars Dale was the reason I began studying education at the University of Oslo. In the spring of 2003, Robert and I represented *Elevorganisasjonen* at the University of Agder's seminar facility, Metochi, in Lesvos, Greece. We spent an entire week discussing research that Erling had undertaken for the Norwegian Directorate for Education and Training (then Læringssenteret), which related to a national upper secondary education professional development programme (Differensierings-prosjektet) on new approaches to adapted teaching. That week defined my choice of career, and, after I enrolled at university,

In university Erling continued to encourage me. I remember my pride when he excitedly showed me that he had cited my first publication in a book he had published. His encouragement stimulated my ambition to take on the task of editing a book while a graduate student in his department in 2007. I am privileged to have enjoyed Erling's support and influence during my time in *Elevorganisasjonen* and as a student. It was a great loss to my research project when, only half a year after its commencement, he passed away. Erling was an astonishingly productive and dedicated scholar, and his many books will continue to stimulate and inspire my intellectual work throughout my career.

**

Lastly, I wish to remember my grandfather, Sverre Therkelsen, after whom I was named. Unfortunately, he did not live to see me complete this project, but I know that he would have been very proud of me. Grandpa has been my inspiration in writing since I was a child. He was a principle at Stavern Folk High School (then Fredtun folkehøyskole) from 1955 to 1970 and throughout many of those years he represented the Norwegian Folk High School in the Norwegian Nynorsk section of the Language Council of Norway (at the time called *Norsk Språknemnd*, now *Språkrådet*).

Grandpa published two collections of poems, one young-adult novel, four booklets of hymns and church songs, one collection of aphorisms as well as four mixed publications consisting short stories and poems. No wonder I wanted to write books too! Grandpa encouraged me through milestones such as publishing my first book chapter at the age of 21. His psalms, novels and aphorisms continue to awe me.

In particular, I loved hearing about his experiences as an exchange student in the Midwest and his adventures in 1945 New York. It was with much pride that I followed in his footsteps, and that journey continues.

My thesis is lovingly dedicated to him.

Sverre Tveit
Oslo, May 30th, 2019

Part I: Extended abstract

1 INTRODUCTION 1

1.1 COMPARATIVE STUDY OF ASSESSMENT POLICIES 2
1.2 THE THEORETICAL BASIS OF THE STUDY 3
1.3 MOTIVATION FOR THE SELECTED CASES OF COMPARISON 4
1.4 RESEARCH QUESTIONS, RESEARCH DESIGN AND DATA 7
1.5 OUTLINE OF THE THESIS 8

2 LITERATURE REVIEW 9

2.1 POLICY BORROWING AND LENDING 9
2.2 CURRICULUM, LEARNING OUTCOMES AND ACCOUNTABILITY 10
2.3 EDUCATION GOVERNANCE, INSPECTION AND JURIDIFICATION 12
2.4 TRANSNATIONAL HISTORY OF EDUCATIONAL ASSESSMENT 13
2.5 NATION STATES' USE OF ILSA IN POLICYMAKING 14
2.6 THE COMPARABILITY OF EDUCATIONAL ASSESSMENTS 16
2.7 FORMATIVE ASSESSMENT 18
2.8 ASSESSMENT *FOR* LEARNING POLICIES AND PROGRAMMES 20
2.9 CONCEPTUAL UNDERSTANDING OF ASSESSMENT PURPOSES 22
2.10 SUMMARY OF THE LITERATURE REVIEW 27

3 THEORETICAL PERSPECTIVES 28

3.1 POLICY LEGITIMATION 28
3.1.1 Legitimacy and legitimation 29
3.1.2 Neo-institutional theories 30
3.1.3 System-reflection theories 32
3.1.4 Policy borrowing and lending 33
3.1.5 Transnational and supranational features of policy legitimation 33
3.1.6 Educational governance and the roles of ILSA 34
3.2 THE LEGITIMACY OF EDUCATIONAL ASSESSMENT POLICIES 36
3.2.1 The validity and reliability of educational assessments 36
3.2.2 The use of educational assessments in educational administration 39
3.2.3 Formative assessment 40
3.3 ANALYTICAL FRAMEWORK FOR RESEARCHING EDUCATIONAL ASSESSMENT POLICIES 41
3.3.1 The legacy of the formative and summative assessment distinction 41
3.3.2 A new distinction between roles of educational assessment 42

4 RESEARCH DESIGN, DATA AND METHODOLOGICAL APPROACH 45

4.1 METHODOLOGICAL APPROACH TO COMPARING THE CASES 46
4.2 RESEARCH DESIGN AND DATA 47

4.3 Policy document analyses 48
4.4 Analyses of expert interviews with policymakers 50
4.5 International policy and research document analyses 52
4.6 Ethical considerations 53
4.7 My role as a researcher 54
4.8 Limitations of the study 55

5 SUMMARY OF THE ARTICLES AND FINDINGS 57

5.1 Article I: Educational assessment in Norway 57
5.2 Article II: New modes of policy legitimation in education 58
5.3 Article III: Ambitious and ambiguous: Shifting purposes of national testing 59
5.4 Article IV: (Trans)national trends and cultures of educational assessment 61
5.5 Overview of the findings related to the research questions 63

6 DISCUSSION 69

6.1 New modes of policy legitimation 70
6.1.1 The role of the OECD in nation states' policymaking 70
6.1.2 Constructions of "world situations" 73
6.1.3 New modes of policy legitimation: A Scandinavian governance turn? 74
6.2 The legitimation of assessment purposes 75
6.2.1 Conflicting purposes of educational assessment 75
6.2.2 Implications of the timing of embracing standardised testing 76
6.2.3 Epistemological differences of subject- and skills-based assessments 77
6.2.4 Disputes over formal grading in Scandinavian education 78
6.3 Borrowing and lending in the formative assessment policy and research discourses 79
6.3.1 Definition problems of the formative and summative distinction 80
6.3.2 The British renaissance of formative assessment discourse 81
6.3.3 The Norwegian borrowing and lending context 83
6.3.4 The diluted meaning of formative assessment 84

7 IMPLICATIONS OF THE STUDY 87

7.1 Theoretical contributions for comparative research 87
7.2 The *professional judgments* versus *external measurement* typology 88
7.3 The Scandinavian legacy of prohibiting formal grading 88
7.4 The legitimacy crisis of the quest for just assessments 89
7.5 The idiosyncratic contexts of formative assessment 89
7.6 A blow to the Nordic model of education? 90
7.7 Exposing problems of the formative and summative assessment distinction 90

8 REFERENCES 91

9 APPENDICES 106

List of Tables

Table 1: Assessment *of* Learning and Assessment *for* Learning 22
Table 2: Categories of purposes for the use of educational assessment judgments..... 25
Table 3: Different conceptual understandings of formative and summative assessment 26
Table 4: A new distinction between the roles of educational assessment 43
Table 5: The sub-studies and data..... 48
Table 6: Policy documents analysed in the study 49
Table 7: Expert interview data 51
Table 8: Three modes of policy legitimation..... 59
Table 9: Three purposes of educational assessment 60
Table 10: Roles and transnational trends of educational assessment 62

List of Figures

Figure 1: Professional (subjective) judgments vs. external (objective) measurement..... 61

List of Appendices

Appendix 1: Policy documents analysed (sub-study III)
Appendix 2: Analyses of policy documents (sub study III)
Appendix 3: Interview guide example from the Norway case
Appendix 4: Interview guide example from the Sweden case
Appendix 5: Ethical approval documentation
Appendix 6: Five types of formative and summative assessment distinctions

1 Introduction

Today, there is a large focus on educational assessment and associated reforms, political debates and disputes in education politics and policymaking. This issue is framed by national and international political discourses that are shaped by the global knowledge economy, with its emphasis on educational outcomes. As such, changes and reforms to educational assessment policies can be perceived as necessary developments that reflect a modernisation process and involve a constant (evolutionary and organic) development. Nation states need to expand and improve their approaches to educational assessment in order to respond to changes in the labour market, developments in terms of new technology and other relevant societal changes.

The legitimation of educational assessment policies also involves several challenges for policymakers that are related to principal features of educational assessment as such. Research studies into formative assessment and Assessment *for* Learning (AfL) programmes have revealed that there is no "royal road" to good assessment practices (Black & Wiliam, 2005). There is also a conflict between nation states' use of assessment for educational administration and the role of the teacher profession, which may feel that these assessments are designed to hold them accountable for outcomes and thus undermines teacher autonomy. Furthermore, research on the validity and reliability of educational assessment has illuminated fundamental problems of comparing students' level of attainment in comparable and just ways (Kane, 2010; Messick, 1989). Thus, policies for educational assessment need to tackle fundamental problems for which there is no ultimate solution. This implies that the legitimacy of educational assessment policies is constantly threatened, which may explain why policy borrowing is widely used in nation states' legitimation of educational assessment policies.

The aim of this study is to investigate how policymakers engage in international research and policy discourses in the legitimation of nation states' educational assessment and testing policies, and to explore the tensions related to the purposes of assessment associated with these policy discourses and legitimation processes. The study's articles report on how Norwegian and Swedish governments have substantiated and justified educational assessment policies and reforms. This extended abstract takes a wider conceptual approach to discuss borrowing and lending in nation states' legitimation of educational assessment policies.

1.1 Comparative study of assessment policies

An advantage of a comparative study of educational assessment policies is that it confronts us with the implicit idiosyncratic conceptual understandings that underpin any single country's policies and practices. To come to terms with and compare countries' policies, we need to develop theoretical perspectives and concepts that are useful for comparing different cultures. As such, this thesis sheds light on educational assessment policies perceived as *phenomenona*. It explores the principal reasons why educational assessment can cause policy legitimation crises, which in turn helps explain the extensive emphasis on transnational trends in order to legitimise national reforms.

Threats to the legitimacy of educational assessment policies are many. Nation states' policymaking and reforms may be responses to political problems and legitimacy crises related to the overall education system. In 2000 the first test of the Programme of International Student Assessment (PISA), facilitated by the Organisation for Economic Co-operation and Development (OECD), was undertaken. In subsequent years, many nation states have faced "PISA shocks", and low test scores in comparison to other countries have been used by politicians to call for or legitimise reforms. New policies and practices for educational assessment are among the solutions proposed to solve these problems.

Any educational assessment needs to deal with the reliability and validity dilemma. There is no ultimate solution for balancing the need to increase the chances that an assessment is accurate and comparable (reliability) with the need to ensure that it assesses what it is supposed to assess (validity). This causes controversies for politicians, policymakers, researchers and educators. Thus, for policymakers, it can be useful to legitimise new policies by referring to international research and other countries' approaches to, for example, testing student achievement.

Tensions between the various uses of educational assessment may cause problems in nation states' policymaking. A short (and not exhaustive) list of what may be associated with educational assessment can give a hint of these tensions. Educational assessment may be used to:

1. select students for various educational programmes and professional careers;
2. provide data to hold teachers, schools and municipalities accountable for outcomes;
3. monitor the quality of nation states' outcomes in comparison to other countries;
4. provide feedback to students to help develop competences and learning strategies;
5. provide feedback to teachers that they can use to improve their instruction.

It is unlikely that these various uses of educational assessment can possibly coexist in a harmonic and undisputed manner. This thesis aims to explore how policymakers engage in international research and policy discourses to tackle legitimation problems related to these tensions between purposes of educational assessment.

Throughout the theoretical and empirical investigations of this PhD project, the distinction between *formative* and *summative assessment* was observed to be widely used by policymakers and researchers alike, both in national and international contexts. It became clear, however, that this conventional way of distinguishing between the purposes of assessment was insufficient for describing research and comparing nation states' policies and practices related to educational assessment. Therefore, to promulgate an analytical framework with novel terms that are useful for analysing and comparing the emphases on different purposes of educational assessment became an important foundation of the study. This framework distinguishes between *three roles of educational assessment*, highlighting the uses of educational assessment to *certify*, *govern* and *support* learning and instruction.

Additionally, historical developments are examined to shed light on how these three purposes of educational assessment have accumulated in the assessment instruments throughout the twentieth century and up to the current date. The study distinguishes between transnational trends of educational assessment related to *meritocracy*, *accountability* and *Assessment for Learning*, to identify key developments. Furthermore, to analyse approaches to the legitimation of educational assessment policies, both historically and in the contemporary setting, the thesis distinguishes between *collaboracy*, *agency* and *consultancy* modes of policy legitimation. These analytical frameworks are outlined in the articles and in the theory chapter of this thesis and serve as theoretical lenses for analysing and discussing the empirical findings.

1.2 The theoretical basis of the study

The theoretical point of departure for this thesis is institutional theory, which considers the processes by which societal structures, including schemes, rules, norms and routines, become established as authoritative guidelines for social behaviour. This perspective helps illuminate how assessment policies gain and sustain legitimacy. *Legitimacy* is defined as the condition of being accepted, plausible and just, while the associated verb *to legitimate* means to justify something or to make something plausible and acceptable (Andersen, 2009). Perceived in the international context of nation states' policymaking, policies can be legitimised by referring to

transnational research and policy discourses and supranational agencies such as the OECD and the European Union (EU).

Classical sociological perspectives on legitimation through externalisation (Berger & Luckmann, 1966; Parsons, 1960; Weber, 1922/1958) are useful for shedding light on these processes. These perspectives can be divided into two strands: neo-institutional theories (DiMaggio & Powell, 1983; Meyer & Rowan, 1977; Weick, 1976) and system-reflexive theories (Luhmann, 1985; Luhmann & Shorr, 1979/2000; Schriewer, 1988, 1999, 2003, 2004, 2014). The former emphasise how globalisation processes lead to an increasingly homogeneous world, while the latter argue that these globalisation processes cause countries to *appear* more similar while in reality the institutional diversity is constantly being deepened.

This thesis discusses both these approaches in order to understand the processes whereby nation states' policies gain and sustain legitimacy. The main focus is on *policy borrowing*, which is the practice of legitimising policies by referring to the policies of other countries (Cowen & Kazamias, 2009; Phillips & Ochs, 2003; Schriewer, 2014; Steiner-Khamsi, 2004, 2010; Steiner-Khamsi & Waldow, 2012). Examples of policy borrowing include pedagogical models of how to undertake formative assessment and national, municipal and school AfL programmes, which are shaped by (national) institutional environments that are difficult to compare (Black & Wiliam, 2005). Given that educational assessment researchers and policymakers claim that these are among the most successful strategies for improving students' and countries' outcomes (Black & Wiliam, 1998a, 1998b), the conceptual underpinnings of these approaches are important areas of research.

1.3 Motivation for the selected cases of comparison

For this thesis, the problems related to the conceptual understanding of educational assessment policies are not just a *theme* to be explored; rather, these issues also pose substantial *methodological* challenges for the comparison of the countries' assessment polices. As such, the study itself is illustrative of the challenges it investigates. This is one of the motives for the comparative approach to study these policy tensions and the two governments' policy legitimation strategies.

Educational assessment policies relate to institutional practices that are distinctive to particular countries' education systems and practices of schooling. The national instruments and formal assessment that underpin meritocratic procedures often rely on "taken for granted" information. This is because "all" members of the national political and practical contexts have

undertaken these tests and examinations and been assessed by teachers in a distinctive national political and practical context that is inherited through generations (Berger & Luckman, 1966). The juridical and political terms are thus highly institutionalised and embedded in the nations' distinct traditions. This implies that many premises of the policies are implicit. Therefore, expert interviews were undertaken at an early stage in order to get an overview of the policy contexts investigated.

The countries that comprise the Scandinavian region share many features in common, including the organisation of the education system and approaches to governing and policymaking. The Scandinavian countries have in common a legacy of high trust in the government and in teachers' capacity to make fair assessments. Compared to, for example, the United States and the United Kingdom, the Scandinavian region has a remarkable trust in government agencies (Marozzy, 2015). It is true that Scandinavian countries have adopted approaches to outcomes-based curriculum steering and the utilisation of assessment standards that are influenced by the English-speaking world. However, the education policies still reflect the Scandinavian legacy of collaboration between the state and the professions, which resist standardisation and privatisation. The educational assessment system reflects a trust in teacher judgments, although this is under threat and increasingly subject to juridification and external control, including school inspections (Hall & Sivesind, 2014; Novak, 2017; Novak & Carlbaum, 2017). Furthermore, while "high-stakes assessment" in the English-speaking research literature most commonly relates to external testing, in the Scandinavian context, the highest stakes with respect to the *certification role* lie in the hands of students' own teachers. In addition, the Scandinavian region's legacy of resisting formal grading in primary education distinguishes it from the education systems of most other countries.

Nevertheless, the region also embeds significant differences. Particularly with respect to national assessment instruments, Sweden's post–World War II approach differs from its Scandinavian counterparts (Lundahl & Tveit, 2014; Ydésen, Ludvigsen, & Lundahl, 2013). As a result of being under Danish rule for almost four centuries (1524–1814), Norway shares its national examination legacy with Denmark. Furthermore, both these countries have similar patterns with respect to psychometric testing, resisting such testing throughout the twentieth century before ultimately, at the turn of the millennium, implementing new national testing programmes.

In the case of Sweden, there is a legacy of national examinations from the seventeenth century up to the second half of the twentieth century. In the 1930s, however, Sweden embarked on a different pattern from its Scandinavian neighbours, with a larger emphasis on psychometric

testing, which ultimately led to the termination of the national examinations in 1968. In the second half of the twentieth century, Sweden developed a distinctive approach to national testing, under greater influence from American scholars than its Scandinavian neighbours.

Sweden's international collaborations through the International Examination Inquiry (IEI) studies in the 1930s and the International Association for the Evaluation of Educational Achievement (IEA) studies in the 1960s and 1970s strengthened the country's expertise in standardised testing. Its termination of the examination system and its emphasis on national tests can be viewed in relation to these international collaborative research efforts. Norway, by contrast, resisted the emphasis on standardised testing that came with these transnational collaborations. In Norway, it was the increased emphasis on accountability in the 1990s and, ultimately, the "PISA shock" in 2001 that prompted national testing. Subsequently, both Norway and Sweden have increasingly emphasised AfL policies, which can also be perceived in the light of the OECD's emphasis on such policies.

Over the past three decades, Sweden has also distinguished itself from its Scandinavian neighbours by taking a substantially different approach to the provision of independent schools (Ringarp, 2011; Wikström, 2005), as well as through its emphasis on and different approaches to school inspection (Hall & Sivesind, 2014), which has implications for our understanding of what is perceived as just assessment (Vogt, 2017; Waldow, 2014). This suggests that despite the global attention both Norway and Sweden have attracted for their successful economic and social development associated with the "Scandinavian welfare state" and "the Nordic model", and despite their similar comprehensive school systems and principles of free and democratic education, there are large differences between the two countries' contemporary educational assessment policies.

These differences both construct and reflect different premises with respect to educational assessment policies and, in turn, diverging premises of policy legitimation. Taken together, these differences in terms of tradition and the contemporary setting helped substantiate the selection of Norway and Sweden as countries for comparison. To better grasp the various premises for the contemporary situation, I decided to explore both selected contemporary assessment reforms (Article I, II and III) and the history of the national assessment instruments (Article IV).

1.4 Research questions, research design and data

Throughout the research process, two overarching research questions guided the study:

RQ1: How do policymakers engage in international research and policy discourses in the legitimation of nation states' educational assessment and testing policies?

RQ2: What tensions related to the purposes of assessment can be identified in nation states' legitimation of educational assessment and testing policies?

To answer these questions, I developed a research design suitable for investigating the policy documents that underpin the most recent reforms of primary and secondary education in Norway and Sweden, as well as conducting expert interviews with policymakers in both countries. The expert interviews included interviews with policymakers at the political and administrative levels of the Norwegian and Swedish ministries of education, as well as the general directors and assessment division directors of the associated executive agencies.

Comparative studies should be undertaken not by relating observable facts but rather by relating relationships or patterns of relationships between institutions (Schriewer, 1988). As mentioned, national tests and examination instruments have different roles in countries' education systems. Thus, it was necessary to develop an analytical framework for comparing the countries' policies. The review of the research literature, the preliminary analyses of the policy documents and the expert interviews with policymakers informed the analytical framework that I promulgated for comparing the two nation states' policies. The study and its four sub-studies are based on three types of data:

- Policy documents for each case
- Expert interviews with policymakers for each case
- International policy and research documents (Eurydice, IEI, IEA and OECD)

The four sub-studies are based on these data in different ways. Sub-study I analyses policy documents from the Norway case to provide a detailed account for the structure of the educational assessment system in Norway. It discusses tensions that can be identified in policymaking during the implementation of the 2006 curriculum reform and the accompanying revision of assessment regulations in 2009. Sub-study II analyses OECD and Eurydice policy documents that were used by policymakers to legitimise the 2014 reform of formal grading age in Sweden. Sub-study III analyses policy documents and policymaker expert interviews for both cases, to compare the emphases on purposes of assessment in the legitimation of the national testing programmes in Norway and Sweden. Sub-study IV analyses international policy and research documents for both cases, related to participation in the IEI, IEA and OECD

studies, as a basis for comparing transnational and supranational influences on the contemporary national assessment instruments in Norway and Sweden.

1.5 Outline of the thesis

Chapter 2 reviews the research literature related to the study's emphasis on policy legitimation and educational assessment policies. Emphasis is given to studies of policy borrowing, education reforms, transnational history of educational assessment, the use of international tests in contemporary policymaking, comparability of educational assessment, formative assessment, and Assessment *for* Learning programmes, as well as conceptual studies of the purposes of educational assessment including various ways researchers distinguish between formative and summative assessment.

Chapter 3 outlines the study's theoretical perspectives related to both education policy legitimation in general and the legitimation of assessment policies in particular. The chapter concludes by promulgating the analytical framework for researching and comparing the roles of educational assessment in different education systems.

Chapter 4 outlines the research design, data and methodological approach to generate and analyse the data, including the approaches to analyse documents and undertake and analyse expert interviews with policymakers. Furthermore, it discusses ethical considerations and limitations of the study. Chapter 5 briefly summarises the findings of the four articles and overviews how each sub-study contributes to the investigation of the study's two research questions.

Chapter 6 discusses the study's findings in response to the research questions in relation to other research studies. Furthermore, based on these findings, the chapter discusses borrowing and lending in the formative assessment policy and research discourses. With the examples of lending in the United Kingdom and the cyclic processes of borrowing and lending related to the reform of the assessment regulations in Norway, it illuminates transnational semantics of educational reform. Conclusively, Chapter 7 identifies the implications of the study and suggests areas of further research.

2 Literature Review

Educational assessment policy is a multifaceted area of study. This literature review overviews Norwegian and Swedish research related to the legitimation of educational assessment policies. In some areas, especially related to the distinction between formative and summative assessment, the review includes studies outside the Norwegian and Swedish contexts that are considered important for these educational assessment and policymaking contexts. The review chapter does not aim to offer a complete or systematic review of studies relevant to the thesis; rather, it reports on studies perceived to be important for the study and studies that demonstrate a need for further research in this field.

First, studies into policymaking and policy borrowing and lending are reviewed. The second section reviews research studies into curriculum reforms, including increased emphasis on learning outcomes and teacher accountability. Third, the chapter reports on studies addressing changes in educational governance. The fourth section review studies on the transnational history of educational assessment. Fifth, it addresses research studies on nation states' use of international tests in contemporary policymaking. The sixth section overview various aspects of the comparability of educational assessments that has been reported in research studies in the two countries. The seventh section includes international research studies on formative assessment. The eight section reviews studies of AfL programmes in Norway and Sweden. The ninth section addresses conceptual studies on the purposes of educational assessment. This section identifies five different ways researchers distinguish between formative and summative assessment. Finally, the findings of the chapter are summarised, with a focus on the perspectives that I have emphasised in the present study.

2.1 Policy borrowing and lending

Research studies have revealed significant developments with respect to the interaction between national actors and external actors in the legitimation of nation states' policies. Waldow (2009) analysed policy discourses in Sweden through two sets of Swedish government committee reports from the 1960s and 1970s to demonstrate the country's legacy of "silent borrowing". Waldow illuminated how Sweden has been largely influenced by the international policy discourse, albeit without acknowledging this in the national policy documents. Furthermore, Waldow found that the Swedish educational research community largely followed the official image of policymaking, with its exclusive focus on the national context. Waldow concluded

that "silent borrowing" was prevalent in Sweden for such a long time because the political culture was "characterised by a powerful myth of rationality and national superiority, favouring strategies of legitimation other than explicit borrowing" (p. 477).

Researching the legitimation of educational policy in Sweden from 1945 to 2014, Ringarp and Waldow (2016) noted that a shift towards an increased utilisation of international arguments occurred in 2007. The authors concluded that this shift related to the declining results on the PISA tests, which changed the perceptions regarding the results of these tests in the public discourse. They concluded that Sweden's "self-confidence as a pioneer country was undermined, which made externalising to world situations more attractive as a legitimatory resource" (p. 6).

Forsberg and Román (2014) observed a shift to a more outcomes-driven curriculum in Sweden with a larger emphasis on measurable outcomes, concluding that "transfers in terms of borrowing and lending have contributed to the formation of [new] conceptions of knowledge" (p. 203). Prøitz (2015a) studied recommendations made by the OECD in three thematic reviews on education policy developments in Norway (in 1988, 2002 and 2011) and the OECD (2013) review of evaluation and assessment frameworks for improving school outcomes (entitled *Synergies for Better Learning*), shedding light on the circular practices of borrowing and lending. Prøitz described a "situation of domestic ideas on a journey uploaded to the OECD and later downloaded as OECD recommendations to support domestic policy directions at the time" (p. 79).

This brief review demonstrates that policy borrowing and lending is significant for the legitimation of the Scandinavian nation states' policies on educational assessment, both as a source of legitimacy and as a space for defining the roles of educational assessment.

2.2 Curriculum, learning outcomes and accountability

An important driver in policymakers' quest to reform educational assessment practices is large-scale assessment. Skedsmo (2011) observed that the implementation of a national system for quality assessment in Norway in 2004 was part of a change from an input-oriented policy to a more output-oriented policy, with an emphasis on learning outcomes, of which national tests are a key component.

The emphasis on tests and outcomes in Norway reflects a general shift from input to output steering that can be observed in many places. "Learning outcomes" is a concept that is increasingly used by nation states when implementing and legitimising reforms. Prøitz (2015b)

demonstrated that the Norwegian curriculum reform of 2006 adopted this concept as a key perspective regarding the role of education, with the outcomes-based curriculum and national tests fostering more of an outcomes-oriented governing of the education system. However, Prøitz observed that the meaning of "learning outcomes" is not necessarily clear. Reviewing journal articles that discuss the concept, Prøitz (2010) observed that the "dominant debate centres on whether learning and the outcomes of learning can and should be stated in full-ended, stable, pre-specified and measurable terms or in open-ended, flexible terms with limited opportunities for measurement" (p. 133). Prøitz (2015b) further observed increased use of the concept "learning outcomes" (læringsutbytte) in Norwegian policy documents between 1997 and 2011, identifying an emerging inside-school focus related to the use of the concept in Norway.

Sundberg and Wahlström (2012) and Wahlström and Sundberg (2015) analysed how international standards-based curricular approaches influenced the recent 2011 school reform in Sweden (*Lgr 11*). They identified a strong policy movement that emphasises standards-based and outcomes-based education systems. In relation to an outcomes-based education that defines educational aims (which Prøitz refers to), a *standards-based* education system further expresses the expected *level* of achievement. These outcomes and standards are formulated by the central government and are expected to be adopted by schools according to a linear top-down model. Holding the teaching profession accountable for results is characteristic of the shift to outcomes-based (Norway) and standards-based (Sweden) Scandinavian education systems, where educational assessment is increasingly used in the governing of education.

Researching how national tests influence teachers' professionalism, Mausethagen (2013a) observed that accountability policies can cause teachers to view tests as external to their work if they feel that the tests undermine rather than support their role in boosting student engagement in the learning process. Mausethagen (2013b) noted that a significant development in Norway over the past years has been that people have become more accountable to scientific knowledge, and that this can be related to the "PISA shock".

Researching teachers' continuing professional development (CPD), Wermke (2013) demonstrated that in comparison to the German teaching profession, the Swedish teaching profession is governed "much more by accountability" (p. 120). Teachers are monitored for efficiency by visibly measurable entities. As it becomes critical to achieve the determined goals, teachers feel obliged to follow the recommendations provided by the state. "The state producing the standards often also provides the material to work with or to achieve the goals and as such

teachers' CPD is then very focused on the state" (Wermke, 2013, p. 120). This type of governing is perceived by the teacher profession to undermine their autonomy.

These changes to the relationship between the state and the teachers echo developments in Norway identified by Mølstad (2015a, 2015b). Exploring state-based curriculum making in Norway and Finland, Mølstad identified different approaches to steering and controlling compulsory education. Mølstad used Hopmann's (2003) distinction between product- and process-controlled education systems. Mølstad argued that while both Norway and Finland belong to the process-controlled tradition, the way the 2006 curriculum reform was implemented implied a new distribution of responsibilities between the national (state) level, municipalities and the profession, which somewhat undermined school and teacher autonomy.

Combined, this brief review demonstrates a development—in Scandinavia and beyond—whereby determination and control of the attainment of learning outcomes form a basis for curriculum steering, which is potentially at odds with teacher autonomy.

2.3 Education governance, inspection and juridification

Shift from government steering to *governance* is a key issue addressed in Scandinavian education policy research. Hall and Sivesind (2014) distinguishes between *governing* seen as "*purposeful efforts to guide, steer, control, or manage* (sectors or facets of) *societies*", while governance refers to "*discursive patterns that emerge from the governing activities of these actors*" (p. 430). Researching school inspections in Norway and Sweden, Hall and Sivesind (2014) observed that "the quality assessment of schools and inspections based on research-based methods shows a stronger emphasis on evaluative modes of governance in the Swedish inspectoral regime" (p. 453). While both countries express the call for purposive, legal-professional modes of governing, evaluative modes were far more strongly stressed in the Swedish case. Norway focused on legal and pragmatic approaches, while Sweden "additionally emphasized professional and expert-defined approaches as well as regulative modes, which potentially intervene into school practice" (p. 454).

Portraying these developments as "juridification of educational spheres", Novak (2018) asserted that "the legitimacy of the postmodern State in the eyes of its citizens can no longer be taken for granted" (p. 11). Novak (2018) argued that the Swedish 2010 Education Act reflect the institutionalization of a juridified school system, and viewed this example of *juridification* as a "strategy of compensatory legitimation" (p. 11).

Taken together these studies on education governance and school inspection shed light on how the juridification of education create new premises for policymaking and the legitimation of educational assessment policies.

2.4 Transnational history of educational assessment

To understand the legitimacy of educational institutions such as tests and examinations, a historical perspective is important. Lawn (2008) edited a volume including several historical investigations of the International Examination Inquiry (IEI), demonstrating how American approaches to standardised testing made an *Atlantic crossing*. The studies of the Norwegian and Swedish cases undertaken by Jarning and Aas (2008) and Lundahl (2008), respectively, demonstrated that the reception of these approaches differed substantially in the two countries. While Sweden embraced new approaches to standardised testing, Norwegian policymakers and researchers were sceptical. These studies shed light on various important historical premises for the legitimation of educational assessment policies in the two countries.

Jarning and Aas (2008) noted that the *Examen Artium* tradition in Norway and Denmark (legislated in 1809) and the *Studenteksamen* in Sweden and Finland (legislated in 1824) are the functional equivalents of the German *Abitur* and the French *Baccalauréat* and thus belong to a pattern of key institutions of continental European education systems (Jarning & Aas, 2008, p. 195). In a study of the Norwegian contribution to the 1930s IEI, Jarning and Aas (2008) observed that the meritocratic procedures were a key focus of the project:

> The focus on fair selection and equality for the talented is represented not least by the meritocratic focus on the control of the validity of marking and comparison of the marking of local teachers and results of the systematic peer assessment of the national examinations. (p. 197)

According to Jarning and Aas (2008), in Norway the "use of testing as a functional alternative or supplement to professional assessment and examination was a contested position already in the 1930s, and not least from the late 1950s" (p. 198). As observed by Lundahl (2008), the Swedish IEI team had a considerably different understanding of the challenges with respect to ensuring fair meritocratic procedures. Fritz Wigforss, a key contributor to the study, called for more use of standardised tests in Sweden. He was convinced that teachers, when equipped with sufficient standardised instruments, were more capable of making comparable judgments than the existing examination system. Wigforss was at the time also involved in a governmental report investigating the prospects of abolishing examination entrance tests and instead allowing elementary school marks to serve as instruments for selection. His position was that "if standardised marks could show better correlation with school success, then entrance tests would

be unnecessary" (Lundahl, 2008, p. 160). Lundahl observed that, in part as a result of Wigforss's influential position in the Swedish IEI research team and the government report, the IEI study's insights regarding new approaches to standardised testing manifested in substantial changes to the meritocratic procedures of post–World War II Sweden.

Lundahl and Waldow (2009) observed that the establishment of the State Psychological and Pedagogical Institute (SPPI) in 1942 was especially critical for Sweden's adoption of (American) standardised tests. The institute was assigned the responsibility of developing new forms of tests that could replace entrance tests and courses related to educational testing. Torsten Husén was one of the instructors for these courses. Lundahl (2006, 2008, 2019) observed how Husén, a professor at the Stockholm Teacher College from 1956 to 1971, exercised a large influence on the Swedish education system for decades. From his position as chair of the IEA from 1962 to 1979, during which time it embarked on several studies in mathematics and science, Husén contributed to the development of a range of new tests used in primary and secondary education in Sweden.

Taken together, these reviewed studies demonstrate that borrowing and lending have shaped the emergence of nation states' assessment instruments. Thus, transnational historical perspectives are necessary to understand the premises of nation states' contemporary policies.

2.5 Nation states' use of ILSA in policymaking

With regards to the contemporary context of international large-scale assessment (ILSA), the PISA tests of the OECD are most known in the public discourse. The PISA tests measure a representative sample of the reading, mathematics and science attainment of the member states' children at the age of 15. Lundgren (2011) pointed to that OECD's role in nation states' policymaking long preceded the PISA tests. Established in 1968, the OECD's Centre for Educational Research and Innovation (CERI) began providing policy recommendations to member countries. Pettersson (2008) addressed that the OECDs emphasis on school and teacher accountability for improving learning outcomes can be related to its mission of economic co-operation and development, and that this explains its significance in the political discourse.

Writing about the history of IEA, former executive director (1997-2014), Wagemaker (2013), contended that "IEA is recognised as the pioneer in the field of cross-national assessment of educational achievement" (p. 13). According to Wagemaker (2013), the aims of the IEA are as follows:

- Provide high-quality data that will increase policymakers' understanding of key school- and non–school based factors that influence teaching and learning;
- Provide high-quality data that will serve as a resource for identifying areas of concern and action, and for preparing and evaluating educational reforms
- Develop and improve the capacity of education systems to engage in national strategies for educational monitoring and improvement;
- Contribute to the development of the worldwide community of researchers in educational evaluation. (pp. 13–14)

Wagemaker (2013) noted that the IEA tests are substantially different from the PISA tests. The distinguishing features of PISA are "its age-based sampling design and its non-curricular, skills-based focus on assessment" (p. 15). Wagemaker (2013) points to that in Germany, a "TIMSS shock" occurred after the release of the Trends in Mathematics and Science Study (TIMSS) results for 1996/1997, which revealed poor mathematics and science outcomes along with social disparity. This was followed by a "PISA shock" when the first PISA results were published in 2001 (Wagemaker, 2013, p. 20). Based on a review of a range of countries, Wagemaker (2013) concluded that there is "compelling evidence that the insights provided by analysis of these [ILSA] data have influenced curricular and instructional reforms" (p. 21).

Kelleghan (2001) observed that, while many Western countries were already utilising the IEA tests to measure students' level of attainment, the 1990s saw an increase in the tests' global influence through their impact on Eastern European and developing countries. Following the declaration of the World Conference on Education for All (UNESCO, 1990), several less-developed countries embarked on national testing programmes, using expertise developed from cross-Atlantic research collaborations.

Grek and Lawn (2009) and Lawn and Grek (2012) observed that "Quick" global and European-level policy comparisons increasingly inform nation states' policymaking. Lundahl and Waldow (2009) identified how "quick languages" (e.g. through comparisons of nation states' outcomes on international tests) frame educational policy discourses and make them accessible to wider circles of participants.

Combined, these reviewed studies demonstrate that ILSAs are important sources to information in nation states' policymaking, both by conditioning the national curriculum reform discourses and by facilitating expertise in how to develop testing instruments.

2.6 The comparability of educational assessments

The comparability of educational assessments is a fundamental issue related to the legitimacy of educational assessment policies. While many countries rely more on external standardised tests, Norway and Sweden have a strong legacy of relying on the capability of teachers to determine students' level of attainment. In any context, this is a complex and difficult matter, especially when judgments are based on written evidence.

To address the legitimacy of, for example, national testing instruments with respect to the comparability of judgments, we should consider the inter-rater reliability of the tests. However, such studies are scarce in Norway and Sweden. One of the few inter-rater reliability studies of examinations in Norway was undertaken by Berge, Evensen, Hertzberg and Wagle (2005). For the Norwegian written examination for lower secondary students, the researchers observed a coefficient of 0.69, which the researchers contended was high compared to other inter-rater reliability studies of written examinations. They argued that the extensive use of guidelines and training for the assessors explained the relatively high level of inter-rater reliability.

While not reporting in terms of correlation coefficients, according to Andresen, Fossum, Rogstad and Smestad (2017) a high level of comparability between the markers was observed for the examination for year-10 mathematics in 2017. Bjørnset, Fossum, Rogstad, Smestad and Talberg (2018) came to the same conclusion for this examination in 2018. As addressed by Brown, Glasswell and Harland (2004), we can expect a higher level of comparability for such tests in comparison to more open formats.

When investigating the research on the reliability and comparability of assessments, it is striking to observe that for many high-stakes assessments, inter-rater reliability is *not* investigated. With the notable exception of the Norwegian language examinations for secondary education (mentioned above) and the recent studies of the year-10 mathematics examination, inter-rater reliability for national written examinations in Norway has not been explored in research studies (DET, 2019). The question of what should be perceived as *sufficient* inter-rater reliability does not appear to be on the Norwegian policy agenda.

Instead, the authorities' and the public's concerns with respect to the reliability of assessments have centred on the comparability between schools' grading practices, measured based on the (diverging) discrepancies between national examination grades and the overall achievement grades assigned by students' own teachers. Investigations into the comparability between Oslo schools' outcomes on national examinations and the final grades determined by

teachers concluded that students in the capital city were not treated equally and that some schools consistently had higher discrepancies than other schools (Kommunerevisjonen, 2009, 2013).

Research studies into teachers' grading practices in Norway are limited. Prøitz (2013) researched differences in grading practices across five school subjects by interviewing 41 teachers in four lower and two upper secondary schools in Norway. The findings suggested that the particular school subject matters when teachers assign final grades to students. For example, in arts and craft, the teachers reported using grading practices relying on a culture of strong assessment communities, with shared standards and universal grading approaches whereby students are primarily rewarded on the basis of their performance and knowledge. By contrast, science and mathematics teachers referred to the calculation of points as being more important for assigning final grades, which they perceived as ensuring fairness and universality in grading (p. 568).

Cliffordson (2004) observed that overall, irrespective of type of school, there has been considerable grade inflation in Sweden since the implementation of a new grading scheme in 1994. Wikström and Wikström (2005) compared the grades that students achieved in upper secondary school with the grades achieved on the Swedish Scholastic Aptitude Tests (SweSAT), finding that "intra-municipal school competition leads to modest levels of grade inflation" (p. 309). Wikström (2005) researched four hypotheses for explaining the increased grade point average in Swedish upper secondary education over the course of six years (1997–2002), concluding that the increase "cannot be explained by better achievements, selection effects or course choices, which means that standards have been lowered" (p. 125). In other words, there has been grade inflation. Wikström (2005) suggested that this was an effect of "leniency in the grading system in combination with pressure for high grading, related to the upper secondary school grades' function as an instrument for selection to higher education" (p. 125).

Due to these problems of grade inflation, the Swedish government mandated the School Inspectorate to re-mark samples of the national tests for schools. Gustafsson and Erickson (2013) called into question the premises of the Swedish School Inspectorate's verification of schools' grading practices in 2011. According to the researchers, the sample was not representative, the inspectors used a different scale from that used by the teachers, and copies sometimes had marginal legibility. Gustafsson and Erickson (2013) concluded that "the results are thus not as clear-cut as suggested by the reports and media releases, which is because a school inspections logic rather than a research logic was applied in designing, conducting, and

reporting the studies" (p. 69). Novak and Carlbaum (2017) investigated the public discourse related to the School Inspectorate's re-marking and demonstrated how the policy for strengthening the inspection measures regarding schools' grading practices was largely legitimised based on public (media) discourse, which in turn was sparked by school inspection reports that did not meet scientific inter-rater reliability standards. Researching grade inflation in Sweden, Vlachos (2018) observed that differences between municipal and independent schools were larger when more reliable tests were used as a reference, and that students in independent schools benefitted from more lenient teacher grading.

Taken together, these reviewed studies demonstrate that different approaches have been used to research the comparability of teacher judgments and of external assessments. They demonstrate that concerns related to grade inflation have been a significant factor especially in Swedish policymaking, which may relate to the liberalisation and marketisation aspects of Swedish education policies.

2.7 Formative assessment

Over the past two decades, *formative assessment* has been a key concept in nation states' educational assessment policies. Research on formative assessment is typically related to feedback practices, although the concept has taken wider forms. Early research studies in relation to formative assessment include the meta-study by Kluger and DeNisi (1996), who analysed research studies that found that praise for task performance appears to be ineffective; instead, feedback is more effective when it builds on experiences from previous tasks and provides information on correct rather than incorrect responses. Furthermore, Deci, Kostner and Ryan (1999) investigated the effects of different types of feedback on motivation, concluding that tangible rewards (e.g. stickers and awards) undermine intrinsic motivation, especially for tasks perceived as interesting. They concluded that extrinsic rewards are typically negative because they "undermine people's taking responsibility for motivating or regulating themselves" (p. 659).

The past two decades, attention to formative assessment was largely sparked by Black and Wiliam's (1998a) comprehensive review article "Assessment and Classroom Learning". The authors reviewed 681 publications determined to be relevant and concluded that "innovations designed to strengthen the frequent feedback that students receive about their learning yield substantial learning gains" (p. 7). Black and Wiliam (1998a) referred to several studies that found that the most effective teachers praise less than average and concluded that

praise draws "attention to self-esteem and away from the task" (p. 49). Based on the observed studies, the authors identified several ways in which teachers can improve formative assessment, including the choice of task, the type of classroom discourse, the way questions are asked and the use of tests. In a brief pamphlet addressing the implications of the review, Black and Wiliam (1998b) contended that "formative assessment experiments produce typical effect sizes of between 0.4 and 0.7: such effect sizes are larger than most of those found for educational interventions" (p. 3). The authors illustrated that an effect size of 0.7 in the most recent TIMSS study at that time (1996) would "raise England from the middle of the 41 countries involved to being one of the top 5" (p. 4). Furthermore, the authors pointed to the negative effects of classroom cultures focused on "rewards, 'gold stars', grades, or class ranking", as pupils then "look for ways to obtain the best marks rather than to improve their learning" (Black & Wiliam, 1998b, p. 6).

Hattie and Timperley (2007) reported on an meta-analysis, concluding that the most effective forms of feedback "provide cues or reinforcement to learners; are in the form of video-, audio-, or computer-assisted instructional feedback; and/or relate to goals" (p. 84). Programmed instruction, praise, punishment and extrinsic rewards were among the feedback types that were least effective for improving attainment. Based on these findings, Hattie and Timperley (2007) undertook a conceptual analysis of feedback and reviewed the evidence related to its impact on learning and achievement.

A few years later, Hattie (2009) published a more comprehensive meta-analysis entitled *Visible Learning* which achieved global impact. This was based on a large range of research studies covering areas considered to influence student learning, including the influence of the student, home, school, curricula, teacher and teaching strategies. Based on these research studies and previous meta-analyses, Hattie reached the overall conclusion that setting challenging learning intentions and being clear about what success means are essential for improving student learning. Thus, Hattie considered *feedback* to be a critical strategy for improving students' learning outcomes. While facing increasing criticism for its statistical methods (e.g. Bergeron & Rivards, 2017), Hattie's work has had a substantial influence during the past decade.

In their review of studies of assessment and learning, Baird, Hopfenbeck, Newton, Stobart and Steen-Utheim (2014), however, concluded that "the effects of formative assessment upon learning have been over-sold by some authors" (p. 6). They perceived this to be "unfortunate because the limited empirical research suggests a modest, but educationally significant, impact on teaching and learning" (p. 6). Correspondingly, Kingston and Nash

(2011) argued that Black and Wiliam's (1998a) publication was treated as a meta-analysis, while it is strictly speaking merely a collection of research studies. In a meta-analysis with strict inclusion criteria, Kingston and Nash (2011) yielded only 13 high-quality studies, which involved 42 effect sizes. The weighted mean effect size across these studies was 0.20, far below Black and Wiliam's (1998) often-cited 0.4–0.7 effect sizes.

Hirsch and Lindberg's (2015) systematic review of research on formative assessment demonstrated that despite studies on formative assessment in compulsory school are few, large meta-studies have contributed "to policy decisions advocating large-scale implementation of formative assessment practices in many countries" (p. 5). The authors problematised empirical studies of formative assessment, pointing out that "the umbrella term formative assessment involves so many and disparate phenomena that it is problematic to speak of *one* overall effect" (p. 5). The authors further expressed a concern that the lack of peer learning among teachers and school leaders has led to "pseudo-formative practices" with only an instrumental understanding of formative assessment.

This brief review demonstrates large emphasis on formative assessment in assessment research. The tendency of the literature has moved from pointing to the advantages of such practices to report on the problems of identifying its impact.

2.8 Assessment *for* Learning policies and programmes

The past decade has seen a large emphasis on Assessment *for* Learning (AfL) in nation states' policymaking. This can be related to the large emphasis on the positive effects of formative assessment addressed above. Reviewing the literature on AfL implementation, Hopfenbeck, Tolo, Florez and Masri (2013) observed a range of reasons why AfL policies have faced implementation problems: teachers' resistance to peer- and self-assessment; teachers' resistance to change in teacher and student roles; a lack of commitment from senior staff; shortcomings in teachers' disciplinary knowledge and assessment skills; a superficial understanding of the approach; busy classrooms; a lack of knowledge of how to put AfL into practice; problems in terms of scaling up; high-stakes testing systems and administrative requirements as obstacles; and students' and teachers' beliefs about assessment (pp. 32–35). Bennett (2011) also points out that one cannot be sure of AfL programme's effects unless adequate definitions of the meaning of formative assessment and AfL is applied.

Hopfenbeck, Petour and Tolo (2015) interviewed stakeholders in charge of the AfL programme in Norway, including ministers of education, members of the Directorate of

Education and Training, and key actors such as municipal leaders, teachers, school leaders and students. They observed that despite successful implementation in some municipalities, the programme had not had any effect on student learning outcomes measured on national tests in reading and mathematics. Gamlem and Smith (2013) interviewed students in lower secondary schools that had participated in the Norwegian AfL programme. Developing a typology for types of feedback, the authors found three types to be typical of classroom interaction (A: rewarding—grade giving—punishing; B: approving—controlling—disapproving; C: specifying attainment—reporting—specifying improvement), while a fourth type (D: constructing achievement—dialogic interaction—constructing the way forward) occurred rarely. The authors suggested that types C and D are most in line with AfL strategies and practices, for which dialogic feedback has been shown to be essential.

Jönsson, Lundahl and Holmgren (2015) reported on a large-scale implementation of AfL in Borås, a medium-sized (approximately 100,000 inhabitants) Swedish municipality, where the implementation of AfL began in 2008. The study indicated that the programme was successful in bringing about a change in how teachers talked about teaching and learning and in changing teachers' pedagogical practice towards AfL. The AfL practices were mostly teacher-centred, with the teachers taking most of the responsibility for the assessment. The researchers observed that this led to a high workload for the teachers and that students may act as passive recipients of authoritative feedback rather than active learners. The OECD's (2005) book, *Formative Assessment: Improving Learning in Secondary Classrooms*, recognised the potential of using assessments for improving learning outcomes. The OECD has emphasised AfL's importance in nation states' educational policies (e.g. OECD, 2013) and has offered to review countries' national AfL programmes (e.g. Hopfenbeck et al., 2013, 2015).

An example of a research article advocating for AfL is a position paper from the European Association for Research on Learning and Instruction (EARLI), where established scholars from several European countries set out "to inform policy makers, educators, and fundraisers about the state-of-the-art, the possibilities, and the needs for innovation in assessment" (Birenbaum et al., 2006). Table 1, gathered from their paper, summarises the main perspectives that motivated the authors' claim that a "paradigm shift" from Assessment *of* Learning to Assessment *for* Learning was needed.

Hayward (2007, 2015) reported on Scotland's Assessment is for Learning (AifL), which was one of the first nationwide AfL programmes. This programme largely served as a model for the implementation of the AfL programme in Norway. Implemented in 2001, the Scottish AifL programme involved teachers, local authority representatives, researchers and

policymakers who met in "learning communities" to "identify and to tackle assessment-related issues" (Hayward, 2015, p. 38). The evaluations of the programme reported highly positive progress with respect to formative assessment. However, the reports also identified tensions in schools between formative aspirations and external accountability demands. Looking back, Hayward (2015) observed that these problems have led to "a misalignment of original ideas and practice over time in Scotland and beyond" (p. 38). Hayward concluded that "prepositions that link assessment to learning—as, for and of—can be useful if they focus attention on different purposes for assessment", but were concerned that these prepositions may "turn into an unreflective mantra drawing attention away from the key construct – *assessment is learning*" (p. 38).

Table 1: Assessment *of* Learning and Assessment *for* Learning

Assessment *of* Learning is:	Assessment *for* Learning is:
• One-dimensional • Summative • Apart from the curriculum but driving the teaching ("teaching for the test") • Inauthentic • Context-independent • Inflexible	• Multi-dimensional • Formative • Integrated into the curriculum • Authentic • Context-embedded • Flexible

Taken together, these reviewed studies demonstrate that many countries have introduced national AfL programmes and that there are multiple implementation challenges associated with these programmes.

2.9 Conceptual understanding of assessment purposes

This section reviews conceptual work in the international literature that is considered most influential for the Norwegian and Swedish contexts of formative assessment and AfL policies.

Through the course of five decades since Scriven (1967) coined the distinction between formative and summative evaluation, the distinction has been elaborated and its meaning expanded in different directions in the areas of both educational evaluation (of curriculum programmes, schools and teachers) and educational assessment (of individual students).

Bloom (1968) and Bloom, Hasting and Madaus (1971) soon put the formative and summative evaluation distinction to use in relation to learning and instruction. In their *Handbook of Formative and Summative Evaluation of Student Learning*, Bloom et al. (1971)

outlined the mastery learning theory, which included an emphasis on the interim testing of students' attainment of learning objects instead of testing merely at the conclusion of programmes. "Formative tests" were to be administrated after the completion of appropriate learning units to identify who had—and who had not—mastered the material. For a student who was able to master the tests, the formative tests would "reinforce the learning and assure him that his present mode of learning and approach to study is adequate", while for a student who lacked mastery of the unit, "the formative test should reveal the particular points of difficulty" (p. 54). Bloom et al. perceived the formative (tests) "to determine the degree of mastery of a given learning task and to pinpoint the part of the task not mastered" (p. 61).

Bloom et al. (1971) further clarified their understanding of formative assessment: "The purpose is not to grade or certify the learner; it is to help both the learner and the teacher focus upon the particular learning necessary for movement towards mastery" (p. 61). They further defined summative evaluation as "directed towards a much more general assessment of the degree to which the larger outcomes have been attained over the entire course or some substantial part of it" (p. 61).

Sadler's (1989) article on formative assessment is widely quoted in the assessment literature. Sadler's (1989) motivation for writing this article was that "many of the principles appropriate to summative assessment are not necessarily transferable to formative assessment" (p. 120). Therefore, he set out to develop a "distinctive conceptualization" for formative assessment. Sadler (1989) defined summative assessment as "reporting at the end of a course of study especially for purposes of certification" (p. 120), while the definition of formative assessment placed more emphasis on the need for more targeted feedback approaches. For Sadler (1989), formative assessment was "concerned with how judgments about the quality of student responses (performances, pieces, or works) can be used to shape and improve the student's competence by short-circuiting the randomness and inefficiency of trial-and-error learning" (p. 120).

In the book, *Testing: Friend or Foe?*, Black (1998) outlined three overall purposes of educational assessment, using the following terminology:

1) Formative—to aid learning;
2) Summative—for review, transfer and certification;
3) Summative—for accountability to the public (p. 34).

Black (1998) recognised that while there are potential tensions between these purposes, there is also a potential for synergies: "Instruments developed and trialled carefully by experts for

certification and accountability exercises can be used by teachers to enrich their own range of questions used for the formative work" (p. 34).

Stobart (2008) used a threefold classification of the purposes of educational assessment that, although listed in a different order, largely corresponds to Black's list of purposes. Stobart (2008) claimed that this classification of purposes is conventional:

1) Selection and certification;
2) Determining and raising standards;
3) Formative assessment—Assessment *for* Learning (p. 24).

Stobart (2008) was careful to emphasise that the purposes are often multiple and that perceiving several purposes as present can offer a helpful perspective, adding that there may be shifts that "bring one into the foreground while another fades into the background" (p. 15). The second point, determining and raising standards, relies on the general assumption "that assessment will signal what has to be learned and the level of understanding and skills needed" (p. 24).

In contemporary policymaking in Norway and Sweden, policy and research on formative assessment commonly refer to Black and Wiliam's (1998a) review article (reviewed above), which focused on formative assessment without conceptually addressing summative assessment. A notable change can be traced from their initial 1998 definition of formative assessment ("all those activities . . . which provide information to be used as feedback") to their 2009 definition where assessment practices are perceived as formative when "evidence about student achievement is elicited, interpreted, and used to make decisions about the next steps in instruction" (Black & Wiliam, 2009, p. 9).

Taras (2005, 2007, 2009) argued that it is possible to extract more principles from Sadler's (1989) understanding of the role of summative assessment in formative assessment than is articulated explicitly in his definitions. Taras took the liberty of explicating an understanding of summative assessment, alleging that it is implicit in Sadler's (1989) much-cited theory of formative assessment and interpreted Scriven's (1967) original paper in light of these perspectives. Taras (2007) observed that the reference to goals and standards is important in Sadler's understanding of formative assessment, even though these are not explicitly expressed in the formal definition quoted above. Instead, Taras quoted another passage from Sadler's paper that expresses the premises for effective formative assessment:

> Stated explicitly, therefore, the learner has to (a) possess a concept of the standard (or goal, or reference level) being aimed for, (b) compare the actual (or current) level of performance with the standard, and (c) engage in appropriate action which leads to some closure of the gap. (Sadler, 1989, p. 121)

Taras contended that a joint feature of Sadler's and Scriven's concepts of formative assessment is that a concept of the goal or standard (or reference level, in Sadler's terminology) is intrinsic

to formative assessment. Taras (2005) contended that Sadler "did not wish to create a dichotomy" and that the power of formative assessment and Assessment *for* Learning "is continuing to be eroded because Scriven's advice has not been heeded and a false separation has been created" between summative and formative assessment (p. 476).

Newton (2007) provided an extensive critical analysis of the theoretical basis for the concepts of summative and formative assessment, addressing similar concerns to Taras (2007) but arriving at a different conclusion. Newton (2007) observed that "when referring to the alleged summative purpose (…) researchers tend simply to use the term as a catch-all expression for categorizing any of a variety of different purposes which are predicated on the use of individual summative assessment judgements" (p. 156). Newton argued that the term "summative" evokes the nature of the assessment *judgment*, namely summing up, and that the purposes for which these summative judgments are used are rarely addressed in texts that address the prospects for formative assessment. Newton (2007) noted that "the use of the assessment judgement appears to be central to the definition of the formative function but is not referred to at all in the definition of the alleged summative function" (p. 157).

Newton (2007) rhetorically asked why the distinction between formative and summative is not grounded in the use to which assessment judgments are put, coming up with a straight-forward answer: "Simply because there is no meaningful distinction to be drawn. The rhetoric appears to distinguish between two conceptually distinct types of use to which results can be put; in fact, it simply foregrounds one particular type, the formative use" (p. 157). Newton (2007) suggested that we use the summative term in relation to a *judgment* and not a purpose. Table 2 gathers 18 different categories of purposes for the use of "educational assessment judgments" (summative assessment) that Newton (2007) identified (pp. 161–162).

Table 2: Categories of purposes for the use of educational assessment judgments

1) Social evaluation uses	7) Guidance uses	13) Resource allocation uses
2) Formative uses	8) Qualification uses	14) Organisational intervention uses
3) Student monitoring uses	9) Selection uses	15) Programme evaluation uses
4) Transfer uses	10) Licensing uses	16) System monitoring uses
5) Placement uses	11) School choice uses	17) Comparability uses
6) Diagnosis uses	12) Institution monitoring uses	18) National accounting uses

To come to terms with the various conceptual understandings of formative and summative assessment, the research literature can be classified based on the different ways of defining

formative and summative assessment and the relationship between the two concepts. Table 3 distinguishes between five different ways of defining formative and summative assessment.

Table 3: Different conceptual understandings of formative and summative assessment[1]

Definitions	Authors
Formative and summative assessment definitions distinguishing between the timing, uses, purposes and roles	Bloom et al., 1971; Sadler, 1989; Scriven, 1967
Formative assessment definitions without (explicit) definitions of summative assessment	Black & Wiliam, 1998a, 2009
Formative assessment and summative assessment definitions that explicitly distinguish between summative assessment used for the certification of individual learners and the evaluation of teachers and schools	Black, 1998; OECD, 2005; Stobart, 2008
Definitions that perceive summative assessment as intrinsic or foundational to the formative assessment process	Taras, 2007
Summative assessment understood as a judgment and not a purpose	Newton, 2007; Scriven, 1967

The reviewed studies of the conceptual understanding of the purposes of assessment demonstrate that there is a variety of interpretations and a lack of consensus with respect to understanding the distinction between formative and summative assessment.

[1] Appendix 6 offers an overview of the quotations that form the basis for the classification of the formative and summative assessment definitions. Scriven (1967) is listed in two definition classifications, as it is interpreted in both ways.

2.10 Summary of the literature review

This review of research studies points to how educational assessment policies are torn between a wide range of purposes, tensions, dilemmas and definition problems. The review first (Chapter 2.1) highlights previous research that brings our attention to how nation states' legitimation of educational assessment policies is conditioned by the international context of educational assessment policy and research. These studies show how other countries and international research and policy agencies are important contexts and premises for nation states' policy legitimation. The research studies addressing the curriculum reform, learning outcomes and accountability (Chapter 2.2) sheds light on how the determination and control of the attainment of learning outcomes form a basis for state governing. Further, Chapter 2.3 sheds light on school inspection and other developments portrayed in terms of education governance and juridification that changes the premises of governments' policy legitimation. The historical studies of educational assessment reviewed in Chapter 2.4 illuminate how transnational policy flows have shaped contemporary testing and examination policies.

Another set of the reviewed studies (Chapters 2.5, 2.6, 2.7 and 2.8) addresses assessment-related issues more specifically, including the emphasis on international tests, the comparability of educational assessment, formative assessment and Assessment *for* Learning. Overall, these studies substantiate many challenges that nation states face in developing and sustaining legitimate educational assessment policies.

The review concludes with a conceptual focus (Chapter 2.9) that informs the present study in two ways. On the one hand, this provides an overview of the origins and emergence of the distinction between formative and summative assessment that is widely used in assessment policy and research discourse. On the other hand, it brings our attention to the vastly different interpretations of this distinction and the lack of consensus and conceptual rigour. This highlights a need for conceptual stringency when researching and comparing educational assessment policies.

The reviewed studies form the basis for outlining the theoretical perspectives of the thesis in the next chapter, which integrates theoretical perspectives on policy borrowing and lending using the concept *policy legitimation* to highlight the nation states' role in setting the agenda both in the national and international policy discourses. Furthermore, the chapter develops theoretical perspectives illuminating conflicting roles of educational assessment to establish that purpose tensions represent constant threats to the legitimation of educational assessment policies.

3 Theoretical Perspectives

This chapter outlines the theoretical perspectives of the study, guided by the two research questions. In the first section, the focus is on the study's first research question: *How do policymakers engage in international research and policy discourses in the legitimation of nation states' educational assessment and testing policies?* This section addresses theoretical perspectives on legitimacy and legitimation, with a focus on the international context of nation states' policymaking. Two main positions for interpreting globalisation and institutional processes are outlined: *neo-institutional theories* and *system-reflection theories*. Furthermore, it addresses transnational, supranational and governance perspectives on policymaking and the emphasis on international large-scale assessments (ILSAs) in the legitimation of nation states' testing policies.

The second section focuses on the second research question: What tensions related to purposes of assessment can be identified in nation states' legitimation of educational assessment and testing policies? This section addresses the issues of validity and reliability, the use of educational assessment in educational administration and the use of educational assessment to improve student learning and teacher instruction (formative assessment).

The third section addresses theoretical conceptualisations of the purposes of educational assessment, criticises the conventional use of the formative and summative assessment distinction, and promulgates a new analytical framework identifying *three roles of educational assessment* to facilitate the study's empirical investigations.

3.1 Policy legitimation

Policy borrowing has received increasing attention in educational research over the past few decades (Cowen & Kazamias, 2009; Schriewer, 2014; Steiner-Khamsi, 2004, 2010; Steiner-Khamsi & Waldow, 2012) in tandem with the growing international policy discourse sparked by international comparative studies of student achievement (Benveniste, 2002; Kamens, 2015; Pettersson, 2008). While policy borrowing, strictly interpreted, refers to when "policy makers in one country seek to employ ideas taken from the experience of another country" (Phillips, 2004, p. 54), the term has expanded to a more general meaning related to how a nation's policy is influenced by other countries. Policy borrowing is essentially a way of ensuring that policies gain or sustain legitimacy. But what is legitimacy and legitimation?

3.1.1 Legitimacy and legitimation

As we enter the fragile theoretical landscape of legitimacy, a fundamental epistemological question should be asked: *What is true knowledge?* While physical phenomena, such as colours, appear to us as objective, they are symbolic structures that were once constructed. Berger and Luckman (1966) discussed how common meaning systems are constructed through on-going social processes of *externalisation*, *objectivation* and *internalisation*. Externalisation is the production of symbolic structures that frame the construction of meaning (i.e. the formulation of terms associated with phenomena). Objectivation occurs when these constructions become products that exist outside the producer(s) as a reality experienced in common with others. *Internalisation* is the process by which the objectified world is "retrojected into consciousness in the course of socialization" (p. 61).

Berger and Luckman (1966) discussed legitimacy within the perspective of the transmission of the social world to a new generation. Legitimation, in their words, serves to make the institutionalised "first-order" objectifications objectively available and subjectively plausible. Based on a historical orientation, they discussed how legitimacy is transferred from one generation to the next: "The problem of legitimation inevitably arises when the objectivation of the (now historic) institutional order are to be transmitted to a new generation" (Berger & Luckman, 1977, p. 92). That is, individuals need to create a (subjective) meaning for inherited objective institutional "realities", whose original meaning is inaccessible in terms of memory. Thus, this meaning needs to be interpreted to them through various legitimating formulas.

> The self-evident character of the institutions can no longer be maintained by means of the individual's own recollection and habitualization. The unity of history and biography is broken. In order to restore it, and thus to make intelligible both aspects of it, there must be "explanations" and justifications of the salient elements of the institutional tradition. Legitimation is this process of "explaining" and justifying. (Berger & Luckman, 1966, p. 92)

If we acknowledge that there is no objective answer to the question *what is true knowledge?* we can then rephrase this question to *who decides what is true knowledge?* The many alternative answers to this question foreshadow the complexity encompassing research into the legitimacy of policies for determining the attainment of educational goals (or standards). We need to look not only at the social actors but also at the social systems and how these relate to the processes and roles of educational assessment. Lundahl (2006) noted that as a result of Berger and Luckman's (1966) perspective, legitimation becomes a process that explains the persistence and safe-transition of institutions.

Many refer to Max Weber's (1922/1958) identification of *three types of legitimate rule* (legal, traditional and charismatic authority) as the most significant contributions to institutional theory (Scott, 2008). Parsons (1960) put Weber's theories to use in modern society by examining the relation between an organisation and its environment, emphasising how individual actors internalise shared norms, which in turn form the basis for their actions. Two schools of comparative education research have evolved from Weber's and Parson's foundational perspectives on legitimation, adopting different perspectives on institutional practices: *neo-institutional theories* and *system-reflection theories*. In the following sections, I discuss these two perspectives to substantiate a theoretical framework for investigating policy legitimation.

3.1.2 Neo-institutional theories

Neo-institutional theory criticises institutional theory (such as that of Parsons) for being overly concerned with stability and order. Meyer and Rowan (1977) embraced the view of institutions as complex networks of cultural rules, albeit focused on how these rules become disconnected from their original meaning. Drawing on Berger and Luckmann (1966), Meyer and Rowan (1977) developed a more critical perspective, addressing how formal structures are manifestations of powerful institutional rules, which become rationalised myths that bind organisations:

> Formal structures are not only creatures of their relational networks in the social organization. In modern societies, the elements of rationalized formal structure are deeply ingrained in, and reflect, widespread understandings of social reality. Many of the positions, policies, programs, and procedures of modern organizations are enforced by public opinion, by the views of important constituents, by knowledge legitimated through the educational system, by social prestige, by the laws, and by the definitions of negligence and prudence used by the courts. (p. 343)

Meyer and Rowan (1977) addressed how institutional myths define new domains of rationalised activity from which formal organisations emerge. Other organisations, in turn, expand their formal structures to become isomorphic (similar or corresponding) to these new myths (p. 345). DiMaggio and Powell (1983) described *institutional isomorphism* as efforts to achieve rationality in situations of uncertainty and constraint, which leads to the homogeneity of organisations' structures (p. 147). They developed a typology of three mechanisms for isomorphic institutional effects: *coercive*, *mimetic* and *normative*.

Coercive isomorphism results from both formal and informal pressure exerted on organisations by other organisations upon which they are dependent, as well as from cultural expectations in the societies within which organisations operate (DiMaggio & Powell, 1983, p. 150). *Mimetic mechanisms* derive from uncertainty (e.g. due to ambiguous goals), which can

make organisations model themselves on other organisations (p. 151). *Normative mechanisms* are associated with the pressure that stems from professionalisation, which DiMaggio and Powell (1983) defined as "the collective struggle of members of an occupation to define the conditions and methods of their work" (p. 152). The isomorphism effects are often present at the same time. For example, a policy securing parents the freedom to choose schools for their children may be justified based on the freedom to choose a hospital for treatment. This legitimation may be associated with constitutional or legal rights (coercive mechanisms). It may however also reflect that principles are borrowed from other institutional contexts (mimic mechanisms) or adhering to professional language of public administration (normative mechanisms).

The three mechanisms detailed by DiMaggio and Powell (1983) help us to understand why actors with responsibility for institutional practices may consider it legitimate to adapt to other institutions' practices and how this shapes the premises of policymaking. Thus, the surrounding national legal context of education policy (coercive mechanisms), policy borrowing from other countries (mimetic mechanisms) and the need of professionals to justify their practices using professional scientific language (normative mechanisms) comprise important aspects of policy legitimation in education.

Inspired by the work of Weick (1976), Spillane and Burch (2006) considered how policy is *coupled*. Policy makers, administrators, and teachers do not simply conform to institutionalized norms and values, "they are instead active agents in the development of the common meaning systems and symbolic processes that build up within and around particular aspects of the technical core" (p. 100). Thus, policymakers identify routines, structures, positions and tools that serve to link different levels of the education system.

These neo institutional perspectives shed light on how common meaning systems are created within education systems. Macro-sociological oriented researchers within the American neo-institutionalist tradition, such as Meyer and Ramirez (2003) and Ramirez, Schofer and Meyer (2018) address these institutional processes on a global level. Researching growing participation in international tests, Ramirez, Schofer and Meyer (2018) situate this development within "a broader world educational culture that favors both a technocratic approach to assessing learning and such progressive educational outcomes as expanded access and broader curricula" (p. 344).

3.1.3 System-reflection theories

System-reflection theories (Luhmann, 1985; Luhmann & Shorr, 1979/2000; Schriewer, 1999, 2003, 2004, 2014) understand the processes of externalisation and isomorphism differently from neo-institutional world culture theories, such as Meyer and Ramirez (2003) and Ramirez, Schofer and Meyer (2018). System-reflection theories emphasise that meaning is developed within the system itself. Legitimation is about developing theories about appropriate action within a system. Opening these theories to their outer environment—externalisation—becomes important in developing "supplementary meaning" and thus achieving or maintaining legitimacy. Schriewer (2004) pointed out that while the neo-institutionalist conception emphasises the global dissemination of ideas, system-reflection theories highlight the *adoptive* mechanisms that operate in varying national reflection contexts (Schriewer, 2004). "Educational system reflection is perceived as an *inevitably culture-specific*, hence *idiosyncratic,* form of theorising and knowledge production" (Schriewer, 2004, p. 489).

Schriewer (1988) discussed how foreign educational systems serve as frames of reference for specifying appropriate reforms for a given nation's education policy (p. 67). Largely influenced by Luhmann (1985), Schriewer's perspective is more concerned with the systems that condition legitimation processes than the actors involved. I draw on his ideas to form a contrast with neo-institutional "world culture" theories that can be perceived as overemphasising isomorphism as an effect of globalisation (cf. Schriewer, 2003, pp. 274–278). While Meyer and Ramirez (2003) portray educational change as "strikingly homogenous and chang[ing] in similar ways around the world" (p. 130), Schriewer foregrounds the implications of the peculiar national context where policy is *borrowed to*. This underlines how it is largely nation states' distinct political and juridical contexts and traditions that condition the framing of the "world situation" and how policies are borrowed.

Drawing on Luhmann and Schorr (1979/2000), Schriewer (1988) identified three types of externalisation to which theories of education can be subject: (1) the scientific nature of the discipline, (2) values and (3) organisations (p. 65). Simplifying the perspectives of Schriewer (1988), we can say that externalisation to world situations can "save" governments from the necessity of relying on values or value-based ideologies. Externalisation to world situations can be an effective strategy for objectifying value-based reasons for decision making in education, and this may be "accomplished in the form of historical descriptions and/or statistical documentations that are recognised as 'scientific'" (Schriewer, 1988, p. 69). As such, references to other countries can make value-based policy implementation more legitimate, as these values

can—through externalisation to world situations—*reappear* as scientific principles that have a higher legitimation potential than values alone.

3.1.4 Policy borrowing and lending

Schriewer (2004) rejected the neo-institutional belief that the world is becoming increasingly similar. Nothing suggests that nationally organised societies will wither away. As a result, we must be prepared for "varying relations between the globalised communication of the sciences on the one hand and, on the other, educational system-reflection's commitment to processing meanings that are deeply rooted in distinctive political and cultural settings" (Schriewer, 2004, p. 532). In other words, while on the surface policies may appear to be moving in the same direction, the differences are in fact constantly being deepened due to their contextual (national) premises. The differences are cemented though distinct (national) levels of discourse, which "again and again opens up possibilities of deliberately selecting alternative externalisations" (Schriewer, 2004, p. 533). Herein lies a paradox with respect to nation states' influence on one another:

> A reflexive context, limited by political boundaries and/or by linguistic links externalises other reflective contexts which, in turn, refer yet to other contexts, with the result that they represent models and possible stimuli to one another. A network of reciprocal references then emerges from this accumulation of observations among nations. This network acquires its own autonomy, which transmits, confirms and accelerates the planetary universalising of reform representations, models, norms, criteria and options. Such a network become an element in the creation of a transnational semantics of pedagogical reform. (Schriewer, 1999, pp. 23–24)

The reflexive context in question in this thesis are the national assessment instruments, national assessment regulations and associated policies. While the international research and policy discourses related to such assessment instruments and regulation uses similar concepts, the distinct national contexts evolve with an expanded depth that cements cultural differences.

According to Schriewer (1988), the national references (e.g. to traditions, beliefs and organisations) that education systems use to legitimise themselves become under threat during times of rapid social, economic and political change. Policy borrowing then "becomes an effective means to radically break with the past through transferring education models, practices, and discourses from other educational systems" (Silova, 2009, p. 299).

3.1.5 Transnational and supranational features of policy legitimation

As mentioned above, Schriewer (1999) used the concept *transnational* to capture the increasingly international form of policymaking in education. While the more common term *international* literally means "between nations", *transnational* means "passing through nations"

(Grek et al., 2009). Dale (2005) defined "supranational" as meaning (literally) "above nations", denoting a "separate, distinct and non-reducible level or scale of activity from the national" (p. 125), referring to how knowledge production is interwoven through a mixture of national and over-national actors that participate in policymaking. According to Dale, "the non-reducibility of 'interventions' or 'policies' to the activities or interests of any particular nation-state" helps distinguish the term *supranational* from *transnational* and *international*, thus indicating "a key element of what is to be understood by globalisation" (p. 125). Characteristic of supranational actors is that the decisions made and policies agreed "are not reducible to, or explicable in terms of, the intentions and interests of individual member states" (Dale, 2005, p. 125).

There does not appear to be a consensus concerning this distinction between transnational and supranational. In this thesis, I have chosen to use the concept *supranational* to describe *formal agencies* such as UNESCO, the OECD and the EU that often (but not necessarily) form the nexus for *transnational* policy flows (Nordin & Sundberg, 2014). Transnational thus refers more broadly to the semantics of policymaking and reform, while supranational refers to the hierarchical dimension of these semantics. Supranational agencies such as the OECD and the EU are generally perceived as exercising influence through the generation and facilitation of policy information that is used to compare nation states. According to Dale (2000), American institutionalists view globalisation as "the presence of a supranational set of ideas, norms, and values that inform—even script—national responses to a range of issues" (p. 436). Dale, by contrast, highlights the local (national) mediation. As such, Dale's criticism of the neo-institutionalists' desire to identify a "world culture" is similar to the critique offered by Schriewer's system-reflection theories discussed above.

3.1.6 Educational governance and the roles of ILSA

Ozga, Dahler-Larsen, Segerholm and Simola (2011) used the term *governance turn* to describe the shift from the practice of policy and administration in its state form (government) to the greater involvement of organisations such as the EU, the World Bank and the OECD. Distinctions between *government* and *governance* are thus commonly used to capture how traditional hierarchical legislative steering becomes influenced by a non-hierarchal *governance* mode of steering, which blurs the distinction between state and civil society (Rhodes, 1997).

Authors give various accounts of these concepts (see recent applications by, e.g., Hall and Sivesind, 2014). Rhodes (2007) emphasises the "governing with and through networks" feature of policymaking (p. 1246). Gunter, Hall and Mills (2014) points to a UK trend in which "non-elected consultants are replacing political debate conducted by publicly accountable

politicians" (p. 519). Lindblad, Pettersson and Popkewitz (2015) observe that global private enterprises such as McKinsey and Company have become increasingly involved in policymaking in recent years. The influence of independent consultants, whether companies or individuals, is also important in understanding how educational governance condition traditional government administration. Grek (2013) notes that policymaking is simultaneously international, transnational, subnational, and national, which implies that global agencies exercise influence on nation states' policymaking in multiple ways.

To understand the legitimacy of nation states' educational assessment and testing policies, it is crucial to understand how ILSA have emerged and are used by nation states in policymaking. Supranational agencies such as the EU and the OECD and international research agencies such as the IEA are used by policymakers to provide synthesised comparative data that can be used in national reform processes. ILSAs are essential in producing and reporting data on nation states' learning outcomes. While public attention seldom reaches below the surface of these outcome comparisons, government officials and politicians dive deeper into the datasets in search of recipes for successful policies. Administration of instruments that produce such comparative datasets is thus associated with power and influence. As Grek (2013) puts it, comparison is not simply informative or reflective: "In fact, it fabricates new realities and hence has become a mode of knowledge production in itself" (p. 698).

Wagemaker (2013) identified the report *A Nation at Risk* by the National Commission of Excellence in Education (1983) to the US as a milestone in the emphasis on standardised testing. The meeting of state governors—prompted by the release of the report—produced "a bipartisan consensus on the need for a statement of national goals for education in the United States" (Wagemaker, 2013, p. 17). This also "tapped into a growing realization across OECD countries that education systems need to operate in a supra-national space, responding to demands to educate citizenry capable of competing in a highly competitive, rapidly changing, globalized social, economic and political world" (p. 17).

Benveniste (2002) described how the emphasis on educational assessment, especially the use of psychometric tests, is part of a global culture that is embraced by education systems worldwide. The increasing emphasis of global agencies on national testing as policy instruments coincided with the IEA launching a new mathematics and science study in 1995 (TIMSS) and a Progress in International Reading Literacy Study (PIRLS) in 2001. Both were now to be undertaken on a cyclic schedule (every fourth and fifth year, respectively), with more emphasis on facilitating the comparison of attainment over time and between countries. It was, however, the OECD's first PISA study (2000) that radically changed the premises of policy

legitimation and education governance across the globe (Meyer & Benavot, 2015), causing a "manic search for best practices" (Kamens, 2015, p. 137). Participation in the IEA studies increased from 12 countries in the first IEA assessment in the 1950s to 79 in TIMSS 2011. Most OECD countries participate in both the OECD and IEA assessments (Wagemaker, 2013, p. 19).

The significant role of ILSAs in nation states' policymaking is important for understanding the legitimacy of the overall education policies and reforms, but also the legitimacy of certain national assessment instruments. In the next main section attention is directed to the principal challenges related to the legitimacy of educational assessment as such.

3.2 The legitimacy of educational assessment policies

While the previous main section addressed how policy legitimation occurs, this main section overviews the specific educational assessment issues that pose challenges to the legitimacy of nation states' policymaking and discusses the most important legitimacy concerns, starting with the impossible task of achieving both valid and reliable assessments.

3.2.1 The validity and reliability of educational assessments

Of paramount importance to the legitimacy of educational assessment policies are fairness in the comparison of student attainment and the appropriateness and representativeness of the assessment instruments and procedures used to determine this. In the field of educational assessment and measurement, these issues are discussed under the headings *reliability* and *validity*, respectively. Messick (1995) noted that "validity, reliability, comparability, and fairness are not just measurement principles; they are social values that have meaning and force whenever evaluative judgments and decisions are made" (p. 5). According to Messick (1989), "validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 13). Defining validity as the "property of inferences that we draw on the basis of assessment outcomes" (p. 127), Wiliam (2008) identified three requirements that must be met for these inferences to have legitimacy:

> The assessment must have adequate reliability, it must address all important aspects of the construct[2] about which we wish to generalize and the responses made by students must not be affected by factors irrelevant to the construct of interest. (p. 128)

In the educational measurement domain, there are multiple theoretical accounts of validity and test validation. Kane (1990, 2006, 2016) contested Messick's *construct validity*[3] with an *argumentation-based approach to test validation* (Newton & Shaw, 2014, p. 136). Although several of these principles can be transferred to settings beyond standardised testing, I follow the lead of Wiliam (2008), who outlined a more basic conceptualisation of reliability and construct validity. In the following, I discuss this in terms of threats to reliability and threats to construct validity.

Threats to reliability concerns the comparability of judgments. According to Wiliam, a reliable test is "one in which the scores that a student gets on different occasions, or with a slightly different set of questions on the tests, or when someone else does the marking, does not change very much" (Wiliam, 2008, p. 128). Wiliam (2008) emphasised three main threats to reliability:

1. Any particular student may perform better or worse depending on the actual questions chosen for the particular administration of the tests;
2. The same students may perform better or worse from day to day;
3. Different markers may give different marks for the same piece of work. (p. 120)

The first and second threats cannot be tackled in single-event, high-stakes assessments with a single or just a few tasks. The only way to tackle this reliability threat is to integrate multiple assessment situations or sources of evidence as a basis for determining students' attainment. This is one reason why teachers' overall judgments for a course of study by many are perceived as the best way to tackle this reliability threat. On the other hand, this approach inevitably implies the use of assessment situations that cannot be controlled by external assessors, which undermines the possibility of inter-reliability checks (the third threat).

The tensions resulting from these threats to the reliability of assessments are tackled in various ways in education systems. Part of the legitimacy of high-stakes external assessments

---

[2] A *construct* can be understood as a measured object (e.g. a defined skill or knowledge that students are expected to persist in or achieve).

[3] Messick (1989) advocated a shift in conceptualisation away from content validity to *construct* validity as the principle aspect of testing. However, more recent psychometric contributions (e.g. Kane 1990, 2006, 2016) have argued that this attempt to unify all aspects of validity in one concept has brought more confusion than clarity (Newton & Shaw, 2014). I have nevertheless chosen to use the concept of construct validity as it is a widely used basic concept in discussions of validity.

(examinations and tests) lies in their *independency* (as opposed to teacher judgments with the risk of subjective bias); however, this approach is typically compromised by the requirement to have multiple assessment situations to facilitate more reliable assessment. Financially and practically, it is impossible to undertake external assessments to the extent that would be required to tackle the threats to reliability listed above. This fundamental dilemma cannot be *solved* in any nation state's policy; thus, it represents a conflict and dilemma that needs to be *tackled* in order for the risks to be minimised and the policies to be perceived as legitimate.

The inter-rater reliability of grades and test scores is commonly reported in terms of reliability coefficients, whereby a correlation of 1 reflects a perfect match while a 0 correlation reflects a complete mismatch. Studies of inter-rater consistency have indicated that whereas tests in science and mathematics subjects can achieve correlations at the higher end (> 0.90), student performance in language and literature subjects—commonly acknowledged to be harder to agree upon—can achieve correlation levels between 0.70 and 0.85 (Brown, Glasswell & Harland, 2004). For the latter types of subjects, professional training in the use of assessment rubrics is essential to improve the correlations (Jonsson & Svingby, 2007).

Threats to construct validity concerns the assessed content. The overall consideration of construct validity is *whether the assessment is relevant to and representative of the assessed domain*. For example, a final examination that intends to determine goal attainment for an entire subject should ideally test all the goals of the respective curriculum. This is virtually impossible, but it is nonetheless important that the test or exam attempts to cover the curriculum in the most representative way possible. Wiliam (2008) defined this validity requirement as construct representativeness: "The extent to which the test adequately represents the constructs we are interested in" (p. 129).

There are two genuine threats to construct representativeness: *construct underrepresentation* and *construct-irrelevant variance*. While we do not want parts of the curriculum to be ignored in the assessment, nor do we want factors other than the (curricular) goals to influence the determination of student attainment. We want the differences in scores to reflect the differences in the capability of the students with respect to the construct (Wiliam, 2008, p. 130).

There is a fundamental tension between construct validity and reliability. For example, standardised tests use a vast number of items in an attempt to maximise reliability. While a large emphasis on multiple-choice items is likely to increase inter-rater reliability, it may lead to construct under-representation because many aspects cannot readily be tested in the multiple-

choice format due to limitations of the format, time and other practical limitations. "Increasing the reliability of a test can therefore result in increasing construct under-representation" (Wiliam, 2008, p. 130). In other words, overemphasis on reliability may undermine validity.

3.2.2 The use of educational assessments in educational administration

In measuring and governing the quality of education, outcomes in terms of academic achievement have come to play an increasing role in recent decades (Hopmann, 2003; Sahlberg, 2016). As a result, to be perceived as legitimate, educational assessment policies are expected to provide information about outcomes that can help governments and local authorities in their administration of the education system. In many countries, the desire of governments for such information on outcomes has led to an increased use of national testing. As the test results are often used to hold governments, municipalities and teachers accountable for outcomes, such policies are often labelled *accountability policies* (Brookhart, 2015).

Sahlberg (2016) included test-based accountability as one of the key features of a global educational reform movement. Hopmann (2003) defined *process*- and *product-control* as two fundamentally different ways of steering the education system through educational assessment. Process-controlled education systems have a national curriculum that provides guidelines to teachers, who are recognised as qualified through national teacher education. In product-controlled education systems, the school sector is divided between private and public providers, with no unified concept of teacher education. Without a licenced teacher profession that determines student outcomes that can be trusted, the legitimacy of outcomes becomes instead reliant on product control. This, Hopmann contends, can explain the increased emphasis on accountability policies in many education systems.

The expansion of countries' national testing programmes, and the increasing emphasis on standardised testing, can be related to developments whereby the product-control dimension of educational administration has become more significant, with an emphasis on the knowledge economy and learning outcomes (Forsberg, 2014; Prøitz, 2010). On the other hand, in education systems traditionally associated with process-controlled educational administration, this emphasis on outcomes has collided with the traditional emphasis on teacher professionalism. The need to legitimate this new emphasis on outcomes may explain the increased emphasis on professional development related to formative assessment, as described in the next section.

3.2.3 Formative assessment

Following Black and Wiliam's (1998a, 1998b) influential review on the effects of formative assessment, many countries have implemented new AfL policies. As reviewed in Chapter 2.7 and 2.8, educational assessment policies are expected to facilitate use of assessment information to improve learning and instruction in line with these perspectives. Why has these perspectives become important in nation state's policy legitimation?

Black and Wiliam (1998a, 1998b) sought to investigate the effects of approaches to educational assessment that support teacher instruction and student learning. They called such assessments "formative assessment", initially defined as "all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities" (Black & Wiliam, 1998a, p. 8). Drawing on these and several other empirical studies, Black and Wiliam (1998a) concluded that the evidence suggested that assessment practices should be changed to the "setting of clear goals, the choice, framing and articulation of appropriate learning tasks, the deployment of these with appropriate pedagogy to evoke feedback . . . and the appropriate interpretation and use of that feedback to guide the learning" (p. 61).

In a subsequent publication, Black and Wiliam (2009) summarised five strategies for improving learning and instruction based on the use of assessment:

- Sharing success criteria with learners;
- Classroom questioning;
- Comment-only marking;
- Peer and self-assessment;
- Formative use of summative tests. (p. 7)

Black and Wiliam's review of formative assessment is commonly used throughout Europe to call for a shift from summative to formative assessment in the policies as well as practices of educational assessment (Kirton, Hallam, Peffers, Robertson, & Stobart, 2007). To further clarify the objectives of formative assessment, the Assessment Reform Group (ARG) in the United Kingdom introduced a set of principles known as Assessment *for* Learning (AfL), aiming to further emphasise the need for assessment practices that support rather than undermine learning and instruction (ARG, 1999). Placing emphasis on the *purpose* of assessment, the ARG listed a range of AfL principles:

- it is embedded in a view of teaching and learning of which it is an essential part;
- it involves sharing learning goals with pupils;
- it aims to help pupils to know and to recognise the standards they are aiming for;
- it involves pupils in self-assessment;
- it provides feedback which leads to pupils recognising their next steps and how to take them;
- it is underpinned by confidence that every student can improve;
- it involves both teacher and pupils reviewing and reflecting on assessment data. (ARG, 1999, p. 7)

Despite it has been embraced in nation states' policies, the formative assessment and AfL literature does not inform us about how these perspectives can be integrated and balanced with other requirements related to education assessment (discussed above). The next section aims to establish an analytical framework that integrates formative assessment with the outlined legitimacy issues related to validity and reliability of educational assessment and the use of assessment in educational administration.

3.3 Analytical framework for researching educational assessment policies

The formative assessment and AfL literature draw on distinctions coined and used to make claims for what kinds of educational assessments are legitimate. Given the rhetorical use of these concepts, together with the inconsistent application by researchers and policymakers (as reviewed in Chapter 2.8 and 2.9), they are not feasible as analytical tools for comparing nation states' educational assessment policies. This section develops Scriven's (1967) original perspectives highlighting the *roles of educational assessment* to promulgate an analytical framework useful for comparing the emphases on purposes of educational assessment.

3.3.1 The legacy of the formative and summative assessment distinction

Scriven (1967) coined the distinction between formative and summative evaluation in the paper *The Methodology of Evaluation*, a philosophical account of research into studies of the effectiveness of school curriculum programmes. Scriven (1967) described evaluation as "gathering and combining performance data with a weighted set of goal scales to yield either comparative or numerical ratings" and distinguished between the roles it was used for. Its role "in the on-going improvement of the curriculum" was called *formative evaluation* (p. 40), while *summative evaluation* referred to programme evaluation's role in enabling administrators to determine the quality of "the entire finished curriculum" (p. 41).

In their *Handbook of Formative and Summative Evaluation of Student Learning*, Bloom et al. (1971) extended the definition of formative evaluation beyond the usage that Scriven (1967) had in mind (Cizek, 2010, p. 5). Bloom et al. (1971) outlined the mastery learning theory, which included an emphasis on the interim testing of students' attainment of learning objects instead of testing merely at the conclusion of the programmes, highlighting that "the purpose is not to grade or certify the learner; it is to help both the learner and the teacher focus upon the

particular learning necessary for movement towards mastery" (p. 61). Bloom's mastery learning programme became highly influential in American education and beyond and can be perceived as foundational for the contemporary understanding of formative assessment in the United States (Cizek, 2010).

Sadler (1989) defined formative assessment utilising Ramaprasad's (1983) definition of feedback: "Feedback is information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way" (p. 4). Policymaking and research on formative assessment in the Scandinavian countries commonly refer to Black and Wiliam's (1998a, 1998b) review. As addressed in Chapter 2.9, Taras (2009) criticised Black and Wiliam (1998, 1998b) for not explicitly relating their theory of formative assessment to the summative side of the distinction from which the concept originates. Nor their developed theory of formative assessment (Black & Wiliam, 2009) discusses the relationship between formative and summative assessment.

Newton (2007) contended that while we tend to classify purposes into smaller numbers of categories, it is more constructive to consider each as a category in its own right. This is not without problems either, as these different uses are likely to be perceived differently in different contexts. A certain level of simplification will always be necessary. However, Newton's list (Table 2) helps make clear the complexity of these matters.

3.3.2 A new distinction between roles of educational assessment

In the development of a new distinction between assessment purposes, I begin with Scriven's (1967) original distinction between formative and summative evaluation and utilise the *role* concept to highlight that different uses can exist simultaneously. I follow Newton's (2007) approach and avoid making a distinction between formative and summative assessment, arguing instead that a summative judgment can be used for a range of purposes (or roles, in Scriven's original terminology).

Integrating the perspectives of Scriven (1967), Sadler (1989), Black (1998), Black and Wiliam (1998a, 1998b, 2009), Newton (2007) and Stobart (2008), I have arrived at three principal *roles of educational assessment* (see Table 4). The main motivation for using three categories is that I follow Black (1998) and Stobart (2009) and recognise that the certification and selection of individual students, on the one hand, and the use of summative judgments to govern the education system, on the other, are fundamentally different uses (see Chapter 2.9).

Table 4: A new distinction between the roles of educational assessment

(I) SCRIVEN'S (1967) ORIGINAL CONCEPTUALISATION OF EDUCATIONAL EVALUATION			
Goal/activity: "Gathering and combining performance data with a weighted set of goal scales to yield either comparative or numerical ratings" (Scriven, 1967, p. 40).			
Formative evaluation "It may have a role in the on-going improvement of the curriculum…" (p. 41).		**Summative evaluation** "In another role, the evaluation process may serve to enable administrators to decide whether the entire finished curriculum . . . represents a sufficiently significant advance on the available alternatives to justify the expense of adoption by a school system" (pp. 41-42).	
(II) SCRIVEN (1967) APPLIED IN EDUCATIONAL ASSESSMENT			
Process	**Goals/standards** To determine educational goal (or standard) attainment		
Role	**Formative assessment** Guide the on-going improvement of learning outcomes	**Summative assessment** Decide and report on the final learning outcomes	
(III) PROPOSED DISTINCTION BETWEEN THREE ROLES OF EDUCATIONAL ASSESSMENT			
Process	To determine educational goal (or standard) attainment		
Role	Used to *support* learning and instruction	Used to *certify* learning and instruction	Used to *govern* learning and instruction
Level	Student and teacher level (classroom assessment)	Student and teacher level (teachers' grading, exit examinations)	Organisational level (schools, municipalities, nation states)
Institutional practice	To identify and communicate gaps between the current and desired attainment levels; used to support learning and instruction strategies	To identify and report the final level of attainment (a grade/mark, examination); used for certification or selection for further education and professional life	To evaluate (aggregated) student attainment data; used to (a) inform decision makers' quality development efforts; and (b) to control application of curricula and regulations

In my framework, I start with Stobart's (2008) first purpose, "selection and certification", which corresponds with what Black (1998) called "summative, for review, transfer and certification". I name this role *certifying* learning and instruction. Stobart's second purpose, "determining and raising standards", overlaps with Black's "summative for accountability to the public". I name this role *governing* learning and instruction. Stobart's third purpose, "formative assessment—assessment *for* learning", is similar to Black's "formative, to aid learning". I name this role *supporting* learning and instruction. All three roles are related to the use of assessment *processes*, defined as *determining educational goal (or standard) attainment*.

An advantage of using this novel terminology for the three roles of educational assessment is that it removes the confusion associated with the formative and summative concepts. As there are fundamental tensions and conflicts associated with the various purposes of educational assessment, a more distinct and clear conceptual approach for coming to terms with the various uses is of key importance for a comparative study into the legitimation of assessment policies.

This framework forms the basis for analysing and discussing tensions associated with the different purposes of educational assessment revealed in this thesis's empirical investigation. It is a fundament for the methodological approach to analyse and compare the data, but as described in the next chapter, it is also a product of the empirical investigations.

4 Research Design, Data and Methodological Approach

The multifaceted landscape of educational assessment policies illuminated in the review chapter (Chapter 2) implies that there is a challenge with respect to establishing the theoretical and empirical constructs of the comparison. Moreover, the theoretical perspectives have revealed that due to the cyclic and reflexive nature of the policy borrowing and lending that characterise contemporary policymaking in many countries, it can be difficult to come to terms with the conceptual underpinnings characteristic of the policy context under study. To investigate how policymakers engage in international research and policy discourses in the legitimation of nation states' educational assessment and testing policies (RQ1) and to explore the associated tensions related to the purposes of assessment (RQ2), it was necessary to develop an analytical framework for researching the specific emphasis on the roles of educational assessment in the two countries' policies.

Theoretical encounters were undertaken to develop functionally equivalent constructs of comparison. The research process included an explorative orientation with respect to the theoretical perspectives and the generation of empirical data. It combined inductive and deductive approaches to analysing the data and the construction of theoretical concepts to support the analyses. The data included a vast set of policy documents for Norway and Sweden, spanning the years 2000–2017, as well as expert interviews with education ministry and executive agency representatives from both countries undertaken in 2013.

Based on the interview data with national experts, and due to new reforms related to grading policies that took place in Sweden in 2014 (after the expert interviews were undertaken), additional data were perceived as important for investigating the research questions. Thus, an additional data collection was undertaken for the Swedish case in 2014, which included an examination of the Eurydice network and how it was used by policymakers in the Swedish education ministry. Furthermore, it became clear that historical perspectives were necessary to understand the contemporary context of national testing in the two countries. Therefore, analyses of participation in the IEI, IEA and OECD studies were undertaken and reported in sub-study IV.

This chapter's first section describes the study's methodological approach to comparing the cases. Section 2 outlines the research design and data. The third section reports on the approach to generating and analysing policy documents, which informed the entire study but

particularly sub-study III. Section 4 reports on the approach to nominating policymakers for interviews and undertaking expert interviews. These interviews informed the overall study, particularly by providing information that helped me search for relevant policy documents for each sub-study. Furthermore, this section reports on the analysis and reporting of the interviews undertaken in sub-study III. The fifth section outlines the approach to analysing the international policy and research documents in sub-study II (Eurydice and OECD data) and sub-study IV (IEA and IEI data). The sixth section discusses the ethical considerations, while section 7 addresses the limitations of the study.

4.1 Methodological approach to comparing the cases

Epstein (1988) distinguished between a relativist and positivist orientation in comparative education research. While positivists use comparison to generalise about schools across cultures, relativists use comparison to grasp the unique character of a nation's education system. The relativist rejects the existence of nomological principles. Rather than being a generalising process, comparison is seen as a method for discovering cultural absolutes (Epstein, 1988). In this study, I aimed to recognise cultural differences but also to develop a sufficiently deep understanding of these cultures to develop and justify theory-based generalisations that could serve as an analytical framework for comparison.

As Bray and Kai (2007) pointed out, "difference and sameness are philosophical opposites, but they are not necessarily antagonistic or mutually exclusive, either in logic or in the real world" (p. 125). In this respect, sameness and difference are not absolute. When conducting a comparison, it is necessary to include both notions. In the logic of comparison, an act of comparing assumes the concept of difference. However, a comparison also includes a search for resemblance between units (Marginson & Mollis, 2001). In this study, I perceived both differences and similarities as important findings that helped to explicate how educational assessment policies, especially the national testing programmes, are legitimised. The development of novel concepts for distinguishing between the roles of educational assessment was useful for identifying sameness and difference. I used an explorative and theory-generating approach to develop theories about the roles of educational assessment, analysing and interpreting the policy documents and expert interview data from Norway and Sweden in the light of the research literature reviewed in Chapter 2 and the theoretical perspectives outlined in Chapter 3.

Analysing text is a hermeneutic act that is riddled with potential sources of misinterpretation. Analysing texts from two different national contexts with different languages adds to the risk of misinterpreting relationships. For this study, one of the strategies for tackling these difficulties in text interpretation and comparison was using the English language as a joint mediator. As described in Chapter 3.3, novel concepts and classifications with respect to the roles of educational assessment were developed to come to terms with the purposes of educational assessment. I went back and forth between inductive and deductive approaches to interpreting the data, ultimately arriving at the analytical framework. Going back and forth between the English definitions and classifications and the policy documents for both countries made it possible to substantiate the classifications of the roles of educational assessment that were emphasised in each country's policies. These categories were in turn used for the empirical analyses. This became foundational for the analysis of the data sets addressed in the next section.

4.2 Research design and data

The study and its four sub-studies are based on three types of data, which are listed below in the chronological order in which the data generation occurred:

- Policy documents for each case
- Expert interviews with policymakers for each case
- International policy and research documents (Eurydice, IEI, IEA and OECD)

The four sub-studies are based on these data in different ways, as shown in Table 5. Sub-study I analyses policy documents from the Norway case to provide a detailed account of the structure of the assessment system and to discuss tensions that can be identified in policymaking during the implementation of the 2006 curriculum reform and the associated revision of assessment regulations in Norway. Sub-study II analyses Eurydice and OECD policy documents that were used by policymakers to legitimise the 2014 reform of grading age in Sweden. Sub-study III analyses policy documents and policymaker expert interviews for both cases to compare the emphasis on the purposes of assessment in the legitimation of the national testing programmes in Norway and Sweden. Sub-study IV analyses international policy and research documents related to participation in the IEI, IEA and OECD studies for both cases as a basis for comparing transnational and supranational influences on the contemporary national assessment instruments in Norway and Sweden.

Table 5: The sub-studies and data

#	Sub-study research question	Data
I	What tensions can be identified in policymaking during the implementation of the 2006 curriculum reform in Norway?	Policy documents (Norway)
II	How did Swedish policymakers use policy descriptions of other countries to legitimise the 2014 reform of the grading age ?	International policy and research documents (Eurydice and OECD)
III	How are the purposes of certifying, governing and supporting learning and instruction emphasised in contemporary written policies and policymakers' legitimation of the national testing programmes in Norway and Sweden?	Policy documents (Norway/Sweden) Expert interviews (Norway/Sweden)
IV	How have transnational and supranational influences shaped the emergence of the contemporary national assessment instruments in Norway and Sweden?	International policy and research documents (IEI, IEA and OECD)

Even though the focus and reporting in sub-studies I and II are limited to one country, the overall study's comparative design informed these sub-studies too. For example, the identification of components of the educational assessment system in Norway in sub-study I benefitted from efforts to come to terms with the corresponding components of the Swedish educational assessment system. Correspondingly, the exploration of grading age policy legitimation in Sweden in sub-study II was partly inspired by policymaking processes and tensions that had been identified in Norway in sub-study I.

4.3 Policy document analyses

Interpreting text using content analysis can be described as a process of searching out underlying themes in the material being analysed (Bryman, 2012). Qualitative Content Analysis (QCA) is a method for "describing the meaning of qualitative material in a systematic way" (Schreier, 2012, p. 1). It is carried out by classifying material as instances of the categories of a coding frame. I used two strategies to generate relevant data: according to *source* and according to *theme* (Schreier, 2012). To determine the relevant sources, I posed the question: *Who is responsible for national educational assessment policymaking in Norway and Sweden?* The answers to this question helped identify the key actors in policymaking. Three arenas of policymaking were identified: the parliaments, the ministries of education and research, and the ministries' executive agencies.

In parliamentary countries such as Norway and Sweden, the parliament forms the highest policymaker level. It decides the laws and broader policies with regard to education. In each country, the Ministry of Education and Research (as it is called in both countries) is responsible for the government's policies, including implementing parliamentary decisions and

detailing regulations based on the laws decided by the parliament. Therefore, policymaking undertaken by the ministries and parliaments was classified as one (ministerial) policy level. Second, both countries have executive agencies responsible for implementing policy and overseeing the national education system. These are called the Norwegian Directorate for Education and Training (Utdanningsdirektoratet) and the Swedish National Agency for Education (Skolverket). A distinction was made between *ministry* and *executive agency* to categorise the two types of government levels and associated policy documents (and policymakers).

The analysis of policies captures the policymaking that shaped the contemporary educational assessment policies in Norway and Sweden between 2000 and 2017. This timeframe was motivated by the policymaking that prepared the implementation of the national testing programme in Norway. This corresponded with the reforms expanding the national testing programme in Sweden. For sub-studies I and III, I identified policy documents through searches on the education ministries' and executive agencies' official websites, searching first for "national tests" ("nasjonale prøver" and "nationella prov" in Norwegian and Swedish, respectively). Within these reports, I then searched for "purpose" ("formål" and "syfte" in Norwegian and Swedish, respectively). Through a second reading of the documents, I paid attention to text that might refer to or be related to the roles of the tests without explicitly addressing their purpose. Third, a snowball sampling approach was used to identify other potential relevant policy documents that were referred to in the policy documents already obtained and analysed. Any documents identified as relevant through the snowball sampling that could not be accessed online were obtained from the respective government body's archives.

Table 6 offers an overview of the policy documents used in sub-study III (see further Appendix 1). All the gathered documents were stored using Nvivo 11 software. The documents related to policy were classified as *policy documents* and assigned the appropriate policy-level categories. They were also classified by country (Norway or Sweden) and year (between 2000 and 2017).

Table 6: Policy documents analysed in the study

COUNTRY	NORWAY	SWEDEN
Ministry of Education	10 policy documents	15 policy documents
Executive agency	3 policy documents	16 policy documents

For sub-study III, the relevant paragraphs were coded as "formål nasjonale prøver" and "syfte nationella prov" ("purpose national tests"). The coded text was transferred to Word documents, where the most central sentences and paragraphs were marked and, in turn, translated into English. The investigation unravelled a comprehensive set of policy document data, especially for the Swedish case, which included 31 policy documents as compared to 13 for Norway.[4] The generated quotations from these policy documents that emphasise the purposes of the national testing programmes are included in Appendix 2. All the quotations in the appendices and in the analysis sections of Article III are English translations from the respective policy documents. Certain limitations to the scope of the analysis of the Swedish documents were necessary due to the vast amount of data.[5]

4.4 Analyses of expert interviews with policymakers

Expert interviews with policymakers were undertaken at an early stage of the study to facilitate an overview of the educational assessment policies and substantiate the searches for relevant policy documents. Sub-study I, however, was undertaken prior to the expert interviews. The entire set of interview data was important for developing the analytical framework and identifying the relevant policy documents for sub-studies II, III and IV, but reporting on the empirical analyses of the interview data was limited to sub-study III. This section reports on the approach to undertaking and analysing the policymaker interviews in relation to both the overall generation of policy information that facilitated the entire study and the specific analyses undertaken in sub-study III.

The nomination of candidates for the expert interviews in this study was different from that of studies whose aim is to target a sample that is representative of the population. The expert interviewees were nominated based on their expertise and responsibilities. For the executive agencies, it was important to interview policymakers with leadership responsibility.

[4] Appendix 2 includes English translations of the quotations from the 13 policy documents for Norway and the 31 policy documents for Sweden. To ensure that all relevant policy documents were included in the analysis, and that each country's history of policy implementation was well accounted for, researchers with expertise in educational measurement and educational assessment policy were consulted in each country. Upon validation, these initial analyses were ultimately reduced to the analysis sections of Article II to meet the requirements of the journal.

[5] In 2008 and 2009, the Swedish Ministry of Education commissioned its executive agency (NAE) and inspection agency (SSI) to control schools' grading practices. These agencies' reports have since been published annually. For this study, only the commissioning letter and first reports were analysed.

I was privileged to have met the director of the Norwegian Directorate for Education and Training and the head of the assessment division during previous academic work. Thus, I was able to recruit both these key experts as interviewees. My Swedish supervisor had previously worked in the Swedish National Agency of Education, and thanks to his recommendations, I was able to recruit directors of the Swedish executive agency who corresponded to the participating experts in the Norway case. This secured comparable high-profile policymaker data sets for the executive agencies. For the ministry level, I wrote to the directors of the relevant ministry departments, asking them to nominate politicians and government officials that could inform my study. As a result, I recruited the state secretary to the minister of education and research in Norway and the political adviser to the minister of education and research in Sweden. Table 7 lists the expert interviews that were undertaken for the study.

Table 7: Expert interview data

Policymakers	**Norway**	**Sweden**
Ministry of education		
Political level	State secretary to the minister of education	Political adviser to the minister of education
Ministry: government official	One government official responsible for assessment policy	Two government officials responsible for assessment policy
Executive agency		
Executive agency: general director	General director of the Norwegian Directorate for Education and Training	General director of the Swedish National Agency for Education
Executive agency: assessment director	Director of Assessment department (Norwegian Directorate for Education and Training)	Director of the Testing and Assessment Unit (Swedish National Agency for Education)

Additionally, I had the privilege of interviewing experts working with national testing instruments in both mathematics and the native language subjects of both countries. Although they were not analysed for this thesis and its articles, these interviews provided contextual insights that were helpful for understanding the "technical core" of national tests and examinations in both countries.

In the preparations for the interviews, I drew on the methodological perspectives of Bogner and Menz (2009) and Littig (2009), combining exploratory and theory-generating expert interviews. Dexter (1969/2006) recommended that the investigator "let the interviewee

introduce to a considerable extent his notion of what he regards as relevant, instead of relying upon the investigator's notions of relevance" (p. 18). Qualitative research interviews may be conducted in an unstructured manner, allowing the interviewee to talk about whatever comes to their mind without the interviewer(s) interrupting in order to retain a rigorous order of questions. Due to substantial differences between the countries' policy contexts and embedded cultural differences, it was considered desirable to prepare a semi-structured interview guide (Kvale & Brinkmann, 2015). Appendices 3 and 4 include examples of the most comprehensive interview guides, which were used for the interviews with the directors of the executive agency assessment department in Norway and the testing and assessment unit in Sweden, respectively. The interviews lasted approximately 45 minutes with the education ministry interviewees and executive agency directors, and 90 minutes with the assessment division interviewees. The interviews gave valuable insights into the ministerial and executive agency processes of policymaking. In the reporting of the interviews, priority was given to the political rather than administrative level of the education ministries.

4.5 International policy and research document analyses

Data for sub-study II was gathered due to new reforms of grading age policies that took place in Sweden in 2014. This included an examination of the Eurydice network and how it was used by policymakers in the Swedish education ministry. I assisted Lundahl, Klapp and Mickwitz (2015) in writing a report on this topic to the Swedish Research Council. This included the identification and analysis of European countries' educational assessment and grading systems as reported in the Eurydice network's web platform in the autumn of 2014.

This investigation was undertaken by compiling information from the Eurydice database based on two research questions: First, *how is the representation of European countries' policies conditioned by the Eurydice database's headings and classifications?* Second, *when searching for the system of formal grading in a country in Eurydice, what kinds of descriptions do the various countries provide?* This information was gathered in rubrics in a MS Word document, which in turn was analysed to identify patterns, as described further in Chapter 5.2 and Article II. It proved to be highly difficult to generate and compare this policy information. As such, the analyses substantiated the problems of comparing educational assessment policies between countries.

The data for sub-study IV included international research documents that reported on Norway and Sweden's participation in the IEI and IEA studies. It was unnecessary to verify

that both Norway and Sweden participate in the OECD PISA studies. For earlier international research collaboration, however, I expected differences to be revealed. I gathered information from the IEA[6] to compare the countries' participation in the IEA studies and was able to establish a significant difference. Furthermore, on site in New York, I investigated International Examinations Inquiry documents located in the Carnegie Collections of the Rare Book and Manuscript Library in Columbia University. I limited these investigations to verifying the two countries' participation and substantiating the impression I had gained from reading two in-depth analyses of the archived documents: *International Examinations Enquiry Committee, Norway, 1929–1937* (Jarning & Aas, 2008) and *International Examinations Enquiry Committee, Sweden, 1929–1937* (Lundahl, 2008).

These analyses suggested that Sweden's participation had substantial implications for developments related to national examinations and testing in Sweden, whereas Norway's participation did not bring about significant developments. I further used these historical findings in relation to the analytical framework distinguishing between the roles of educational assessment that I had developed as a foundation for the entire study. In Article IV, I expanded this framework by relating the two countries' historical developments to the emphasis on transnational trends of educational assessment.

4.6 Ethical considerations

This research project was approved by the Norwegian Centre for Research Data (NSD; see Appendix 5) and followed the general ethical guidelines required for social research in Norway. Section 4.4 describes how I contacted the ministries and executive agencies in the preparations for the study. A dilemma in the expert interviews is that the validity of the data (in terms of the experts' knowledge and responsibilities) relies on the interviewees' formal roles and expertise (Bogner & Menz, 2009; Dexter, 1969/2006). Reporting this information, however, is at odds with the general confidentiality principles of social research (Bryman, 2012). However, it would not have been possible to undertake this study without reporting the formal roles of the interviewees. This means that people that are working in the field of education policy and educational assessment may be able to identify the interviewees.

[6] Note that the web page with this information that I accessed on April 17, 2016 (http://www.iea.nl/brief_history.html), is no longer available.

I chose to tackle this dilemma by informing the interviewees about the risk of indirect identification. Before undertaking the interviews, I made the interviewees aware that even though they would not be quoted by name, the expert interview method required openness about the respective interviewee's formal role in policymaking. This was not a problem for the interviewees, who consented to be referred to by their official title. In their positions of leadership responsibility, it is part of their job to represent the level of government in which they work. From a methodological point of view, however, this situation implies a limitation of the study. The interviewees may have felt that there was a risk that their leaders might sanction (perceived) inappropriate or controversial comments and answers to my questions. Thus, it is possible that my interviewees did not report important relevant information that could have informed the study. Nonetheless, I perceived it as my ethical responsibility as a researcher to make the interviewees aware of the possibility of indirect identification, despite the limitations to the study that this may have caused.

4.7 My role as a researcher

Research ethics is not merely about access to, storage of and use of the collected data. As Kvale and Brinkmann (2015) pointed out, researchers cannot avoid their own background affecting the inferences and choices made. Researchers therefore require reflexivity, which involves reflection upon how we construct social phenomena and our role as researchers in the production of knowledge. According to Bryman (2012), "values intrude in all phases of the research processes—from the choice of research area to the formulation of conclusions" (p. 149). It is thus of pivotal importance that researchers are aware of their own presumptions and open to other perspectives and interpretations.

As a way to deal with this, Kvale and Brinkmann (2015) pointed to the importance of communicative validity with respect to data analyses. Researchers are part of a research community and need feedback from their colleagues to improve their skills in terms of ethical interpretations and the judgments that underpin inferences and conclusions. "If one aims to improve one's skills with respect to ethical considerations, judgments and thinking, one needs feedback from others" (Kvale & Brinkmann, 2015, p. 113, my translation). In addition to feedback from my supervisors, as a member of the National Graduate School of Educational Research (NATED), I received feedback and input on my research throughout the project. In addition, by presenting my work at international research conferences and submitting my

articles to research journals for review, my interpretations, perspectives and judgments were constantly challenged.

My role as a researcher is shaped by my activist background as a student union representative from 1999 to 2003. As a high-school student, I was concerned about the comparability of teachers' grading and called into question the legitimacy of the ways in which the examination system determined students' subject knowledge and skills. Challenged by various theoretical and empirical accounts, this normative perspective has matured for me, but it continues to shape my attitudes and perhaps limit my capacity to take in other perspectives. On the other hand, this background has given me valuable insights into policymaking processes. I early understood the significance of the Scandinavian legacy of involving experts, professional organisations and stakeholders in reform processes to sustain the legitimacy of education policies in general and educational assessment in particular. Throughout the research process, I have aimed to acknowledge how my student activist background shapes my perspectives, seeking to make this an advantage rather than a limitation.

4.8 Limitations of the study

For this study, it was necessary to establish a firm limitation on the empirical data in order to facilitate a theoretical and empirical orientation of the desired depth. I considered that researching policymaking from the national policymakers' perspective was the most plausible point of departure. However, the teacher profession, teacher unions, school leaders and municipalities are other actors that are important for the legitimacy of assessment policies. I acknowledge that including such groups would have added different perspectives to the study.

Another limitation is that I do not know the policy context equally well in the two countries. Therefore, while sub-studies III and IV use an equally comprehensive approach to investigate and analyse both cases' national tests and examination instruments. I placed a deliberate emphasis on the wider curriculum reform context in Norway (sub-study I), while limiting the analyses in the case of Sweden to the reform of formal grading age (sub-study II).

In Norway, the national testing programme was implemented in 2004 and subsequently revised. This act of policymaking, and the associated shifting emphasis on the roles of educational assessment, can be traced in ministry reports to the parliament. In Sweden, by contrast, the national testing programme was already in place. This means that we would expect the *volume* of policy documents to differ in each case, reflecting these different historical premises. Therefore, in order to emphasise the roles of educational assessment in the policy

documents, I focused the investigations not on numerical comparisons (e.g. word frequency analyses) but rather on the qualitative analysis and categorisation of the content.

Quantitative content analyses could, however, potentially shed new light on the qualitative analyses. For example, it would be interesting to undertake word frequency analyses to illuminate and compare the emphasis on formative assessment and AfL in policy documents in Norway and Sweden.

5 Summary of the Articles and Findings

This chapter summarises each article. It concludes by relating the findings to the thesis's two overall research questions in advance of the discussion that follows (Chapter 6).

5.1 Article I: Educational assessment in Norway

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

This sub-study offers an overview of the educational assessment procedures in Norwegian primary and secondary education and describes the reform processes that were initiated in 2002 in response to the "PISA shock". The reform brought the implementation of new national tests in 2004, which underwent considerable changes due to scientific weaknesses and controversies regarding the publication of school results. In 2006, a curriculum reform was implemented with a new outcomes-based national curriculum that included new approaches to using criteria and standards and the integration of cross-disciplinary basic skills. For reading, numeracy and science skills, corresponding national tests were implemented. The development of competence aims in the subject curricula was based on the OECD's Definition and Selection of Competences (DeSeCo) framework. Furthermore, following the reform, a major revision of assessment regulations was undertaken in 2009. Comprehensive government projects and a national Assessment *for* Learning programme, that promoted formative assessment, was implemented. As the article identifies, the current assessment regulations state that students learn best when they:

> understand what they are to learn and what is expected of them; receive feedback that tells them about the quality of their work or performance; receive advice on how they can improve; are involved in their own learning by, for example, assessing their own work and development. (Article I, p. 226)

These regulations correspond almost literally to the four principles developed by Black and Wiliam (1998a, 1998b), which were discussed in Chapter 2.8 and 3.2.3. As such, the study demonstrates that policy borrowing from the international policy and research discourses largely informed the revision of the assessment regulations in Norway. The study also identifies tensions between the state and the teacher profession regarding the implementation of national assessment criteria in the curricula. Furthermore, it discusses tensions related to the purposes of the national tests and the Scandinavian legacy of prohibiting formal grading in primary education.

5.2 Article II: New modes of policy legitimation in education

Tveit, S., & Lundahl, C. (2018). New modes of policy legitimation in education: (Mis)using comparative data to effectuate assessment reform. *European Educational Research Journal, 17*(5), 631–655. https://doi.org/10.1177%2F1474904117728846

Identifying *three modes of policy legitimation* in education, which are illustrated by shifts in Swedish educational assessment and grading policies over the past decades, this sub-study demonstrates significant trends with regard to national governments' policymaking and policy borrowing. It portrays a shift away from *collaboracy*[7]—defined as policy legitimation located in partnerships with and networks of stakeholders, researchers and other experts—towards a greater use of supranational agencies (described as *agency*), such as the OECD, the EU and associated networks, as well as the use of individual consultants and private enterprises (described as *consultancy*) to legitimate policy changes. The framework, outlined in Table 8, integrates Weberian perspectives on traditional authority with the neo-institutional and system-reflexive theories outlined in Chapter 3.1 in relation to (a) the type of actors, (b) their type of authority and (c) the type of institutional processes (mimic isomorphism) that produce the legitimacy. The three modes of policy legitimation in this typology can occur to various extents, both independently and simultaneously.

The article identifies shifts in Swedish educational assessment and grading policies over the past decades, including the controversial change of grading age from Year 8 to Year 6, and the further attempted change to Year 4. It analyses how these changes were legitimised through the nomination of a neuroscience professor who proposed policy changes to the government, allegedly based on OECD reports and policy information presented in the EU-affiliated Eurydice database. This is used as an example to illustrate the identified theoretical perspectives on the modes of policy legitimation. Analysing Eurydice data for assessment and grading policies, the article discusses the functional equivalence of grading policies and validity

[7] *Collaboracy* is a twist on the term *collaborative* and is constructed to operate in tandem with the *agency* and *consultancy* modes of policy legitimation. It might well be described as *collaborative policy legitimation* in other contexts.

problems related to the comparison of such policy information. The analyses were undertaken to *substantiate variations* rather than facilitate comparison.

Table 8: Three modes of policy legitimation

Modes of policy legitimation	Collaboracy	Agency	Consultancy
A) Actors	The government produces legitimacy by nominating stakeholders, researchers and other experts to review and recommend policy changes	The government produces legitimacy by cooperating with formal agencies which *fund*, *commission*, *synthesise*, *review* or *recommend* policy changes	The government produces legitimacy by nominating individual experts or private enterprises to review and recommend policy changes
B) Type of authority	Representativeness and expertise as perceived by the public and professionals	Hierarchical (e.g. supranational towards national; national towards local)	Expertise in line with the targeted policy measures and promoted knowledge basis as defined by the government
C) Institutional processes	Mimic isomorphism, not coercive. Inherent legitimacy maintained through tradition	Mimic isomorphism; sometimes object to coercive isomorphism	Selective mimic isomorphism (selective modelling)

The article illuminates how difficult—or outright impossible—it is to arrive at a comparable notion of grading age and, as such, falsifies a foundational premise of the Swedish government's legitimation of the assessment reform.

5.3 Article III: Ambitious and ambiguous: Shifting purposes of national testing

Tveit, S. (2018b). Ambitious and ambiguous: Shifting purposes of national testing in the legitimation of assessment policies in Norway and Sweden (2000–2017). *Assessment in Education: Principles, Policy & Practice*, 25(3), 327–350. https://doi.org/10.1080/0969594X.2017.1421522

This sub-study promulgates an analytical framework for researching *the roles* of *educational assessment* emphasised in the two governments' assessment policies. This framework serves as a basis for comparing the assessment purpose emphases that shaped the contemporary policies of national testing in primary and lower secondary education in Norway and Sweden in the new millennium. The article outlines a distinction between three purposes of educational assessment (Table 9). Reporting on analyses of policy documents and expert interviews with government

officials from the education ministry and associated executive agency in each country, the article illuminates the (shifting) emphasis on the use of national tests to *certify*, *govern* and *support* learning and instruction.

Table 9: Three purposes of educational assessment

Process	To determine educational goal (or standard) attainment		
Purpose	To *certify* learning and instruction	To *govern* learning and instruction	To *support* learning and instruction
Level	Student and teacher level (teachers' grading, exit examinations)	Organisational level (schools, municipalities, nation states)	Student and teacher level (classroom assessment)
Institutional practice	To identify and report the final level of attainment (a grade/mark, examination); used for certification or selection for further education and professional life	To evaluate (aggregated) student attainment data; used to (a) inform decision makers' quality development efforts; and (b) to control application of curricula and regulations	To identify and communicate gaps between the current and desired attainment levels; used to support learning and instruction strategies

The empirical data included 13 policy documents for Norway and 31 for Sweden. Quotations related to the purpose of educational assessment in the national tests were coded and translated into English. For each country, a comprehensive analysis of the policy documents was undertaken for the investigated period (2000–2017). Additionally, analyses of expert interviews with policymakers in the education ministries and executive agencies (undertaken in 2013) substantiated the analyses. The article demonstrates that both countries struggled to integrate formative assessment into national testing programmes primarily designed to serve conventional governing and certifying roles. It illuminates the governments' ambiguous conceptions of the purposes of assessment, suggesting that this can be explained by the (overly) ambitious political demands for integrating multiple purposes into single testing programmes. The tensions between the certifying, governing and/or supporting roles of the national tests are discussed. In addition, the article discusses how the premises of the policy legitimation of the national tests have changed in Sweden. It relates this development to the tests' more significant role in governing education as a result of the decentralisation of the education system, the liberalisation of independent school policies and the juridification of the education system at large (including a strengthened system of school inspection).

5.4 Article IV: (Trans)national trends and cultures of educational assessment

Tveit, S. (2018a). Transnational trends and cultures of educational assessment: Reception and resistance of national testing in Norway and Sweden during the twentieth century. In C. Alarcon & M. Lawn (Eds.), *Assessment cultures*. Studia Educationis Historica. Berlin, Germany: Peter Lang. https://doi.org/10.3726/978-3-653-06867-2

This sub-study outlines analytical frameworks for identifying the transnational and supranational influences on Norway's and Sweden's national assessment instruments. The study integrates Hopmann's (2003) perspectives on process- and product-controlled education systems with Carson's (2006) comparison of the concept of merit in the French and American republics during the twentieth century. These perspectives help describe the emergence of two distinctly different approaches to educational assessment in the continental European countries and the United States: the *examination culture* (with an emphasis on professional [subjective] judgments) and the *testing culture* (with an emphasis on external [objective] measurements), as shown in Figure 1 (Article IV, p. 140).

Curriculum steering	**Process-control**	**Product-control**
Premises for controlling the curriculum and teachers:	The national curriculum provides guidelines to teachers, who are recognised as qualified through national teacher education.	The school sector is divided between private and public providers, with no unified concept of teacher education. Thus, the emphasis is on external product control instead.

Assessment instruments	**EXAMINATIONS:**	**TESTS:**
Assessment instruments used to govern the education system and its certification procedures rely on:	**Professional (subjective) judgement:** Members of the profession control each other's assessments to facilitate the validity and comparability of assessments.	**External (objective) measurement:** Standardised tests developed by measurement experts facilitate the validity and comparability of assessments.

Figure 1: Professional (subjective) judgments vs. external (objective) measurement

Based on these typologies, the study details the reception of and resistance to American approaches to standardised testing during the twentieth century in Sweden and Norway. The differences in the reception of standardised testing are portrayed through an analytical framework identifying *three transnational trends of educational assessment* (meritocracy, accountability and Assessment *for* Learning) that have been emphasised by nation states in the legitimation of educational assessment policies in general and national testing policies in particular, as addressed in Table 10 (gathered from Table 1 in Article IV, p. 138).

Table 10: Roles and transnational trends of educational assessment

Process	To determine educational goal (or standard) attainment		
Role	To *certify* learning and instruction	To *govern* learning and instruction	To *support* learning and instruction
Trends	Meritocracy (1930s→)	Accountability (1990→)	Assessment *for* Learning (2000→)
Transnational research projects	International Examinations Inquiry (IEI), 1933–1938	**IEA TIMSS**: 1995, 1999, 2003, 2007, 2011, 2015 **IEA PIRLS**: 2001, 2006, 2011, 2016 **OECD PISA**: 2000, 2003, 2006, 2009, 2012, 2015, 2018	OECD, 2005 OECD, 2013

The *meritocracy trend*, focusing on fair certification and selection procedures for individual students, was emphasised in international research projects such as the IEI of the 1930s. The *accountability trend* places more emphasis on the governing of education systems and supranational agencies' role in global competition among nation states. Examples of this include the TIMSS and PIRLS tests (facilitated by the IEA from the 1990s onwards), the PISA tests (facilitated by the OECD from 2000 onwards) and the OECD's role in reviewing and recommending policies. The Assessment *for* Learning trend, emphasising the role of assessment instruments and procedures in supporting student learning, emerged at the turn of the millennium following the publication of research reviews that identified "formative assessment" and the "formative use of summative tests" as successful strategies for enhancing students' and countries' achievements.

The study demonstrates how countries that began with an examination culture have—to a greater or lesser extent—been influenced by the testing culture, which can in part be related to engagement with the transnational trends throughout the twentieth century. Whereas the meritocracy trend from the 1930s influenced the emphasis on psychometric approaches to

educational assessment in Sweden (and ultimately the replacement of examinations with psychometric tests in 1968), it was during the accountability trend from the 1990s onwards that psychometric approaches to educational assessment broke through in Norway (especially as a result of the "PISA shock" in 2001).

These developments are related to the *three roles of educational assessment* in contemporary assessment policies, namely the use of national tests to *certify*, *govern* and *support* learning and instruction. Norway currently has two national assessment instruments that are assigned different primary roles: The examinations primarily have a certifying role, while the national tests primarily have a governing role. Sweden, however, simply expanded its existing national testing programme when the accountability trend brought an increased emphasis on psychometric testing in the 1990s. At the turn of the millennium, the transnational emphasis on Assessment *for* Learning contributed to a new emphasis on the role of assessment instruments in supporting students' learning and teachers' instruction. Both countries "added" this role to their respective national tests.

The study demonstrates that it is essential to acknowledge the different timings of the implementation of the national tests to understand the different cultures of educational assessment. The national tests in Sweden are subject- and disciplinary-based in correspondence with the IEI research project that shaped the meritocracy trend. This reflects how they are used to certify (subject) learning and instruction. In Norway, the national tests are interdisciplinary and skills-based, which reflects the emphasis on skills in PISA, the most influential ILSA programme associated with the accountability trend. Thus, while they are called "national tests" in both Norway and Sweden, these national assessment instruments have undergone different transnational influences characteristic of the transnational trend at the time of implementation. As a result, they serve different roles in contemporary policies.

5.5 Overview of the findings related to the research questions

Across the four studies, several overall observations can be made related to the first research question: How do policymakers engage in international research and policy discourses in the legitimation of nation states' educational assessment and testing policies?

Article IV demonstrates how international research projects and policy discourses (e.g. the IEI study of the 1930s, the studies by the IEA from the late 1950s onwards and the OECD's PISA study from the turn of the millennium onwards) have influenced nation states' educational

assessment policies. The three transnational trends of educational assessment identified in Article IV shed light on how different roles were emphasised as the national and ILSA instruments were developed throughout the twentieth century (and up to today). The article demonstrates that while meritocratic procedures were a key focus in the international research and policy discourses of the 1930s, the focus shifted in the 1990s to a greater emphasis on accountability policies, while Assessment *for* Learning became important for policy legitimation at the turn of the millennium.

The analytical framework of Article II, which distinguishes between collaboracy, agency and consultancy modes of policy legitimation, sheds light on how the legitimacy of educational assessment policies is related to the different ways in which national policymakers engage in the international policy and research discourses to legitimise national policies. Furthermore, the framework helps point out how new approaches to policy legitimation have emerged, with more emphasis on supranational agencies and the use of individual consultants and private enterprises to legitimate reforms. In what follows, this framework is used to report on findings across all four articles related to how policymakers engage in international research and policy discourses in the legitimation of educational assessment and testing policies.

The collaboracy mode of policy legitimation has a long-standing tradition in Norway and Sweden. This is reflected in all the sub-studies. Before the Norwegian or Swedish government draws up a legislative proposal for a new policy, it may choose to appoint a special expert (officially known as a one-man committee of inquiry) or an expert group (a commission of inquiry) to investigate the issues in question. Reporting on matters in accordance with a set of instructions laid down by the government, these operate independently and may include or co-opt stakeholders, researchers, other experts, public officials and politicians. The reports are published in the Norwegian Government Official Report series (Norges Offentlige Utredninger, NOU) or the Swedish Government Official Report series (Statens Offentliga Utredningar, SOU). After a committee has submitted its report to the responsible ministry, it is sent to relevant authorities, stakeholders and the public for consideration. These are given an opportunity to express their views before the government formulates and presents a legislative and/or budget proposal to the parliament. As such, the reforms undertaken are prepared by everyone who is part of the education system.

Articles I, III and IV demonstrate that the implementation of national tests in Norway was based on recommendations from a committee appointed by the education ministry to review new policies. This committee had democratic representation, combining national expertise with professional and stakeholder interests. The international discourses and

arguments are highly present in such committees, albeit (democratically) the national expert, professional and stakeholder interests are in charge rather than the government alone.

Article II points to the strong legacy of the collaboracy mode of policy legitimation in Sweden. The almost decade-long committee review of grading policy in the 1970s is a prominent example of the Swedish government's collaboration with experts, the profession and stakeholders when reviewing and formulating new policies. Article II demonstrates that even further back, starting in the 1930s, key scholars in Sweden, such as Wigforss, Husén and other professionals, contributed to collaborative research projects where new methods for assessing student attainment were developed. Article III sheds light on the ongoing revision of the national testing programme, which also reflects the government's collaboration with experts, the profession and stakeholders. As such, the new modes of policy legitimation outlined below supplement rather than replace the collaborative tradition.

The agency mode of policy legitimation can also be observed across all the sub-studies. Articles I and III demonstrate that the implementation of the Norwegian national testing programme was modelled on the OECD PISA tests. Following the "PISA shock" in December 2001, the government expanded the mandate (and number of members) of a commission that had been appointed to evaluate and propose reforms for the primary and secondary education systems. Article I demonstrates that the accountability trend, in which the OECD and the PISA tests play a substantial role, was a driving factor in Norway's implementation of standardised national tests. Providing a broader account of the development of the educational assessment policies in Norway, Article I also demonstrates the supranational influences on the entire education reform of 2006. The resulting outcomes-based national curriculum can be seen as an example of how the OECD's increased emphasis on learning outcomes has been taken up in national reforms. Furthermore, Article I demonstrates that the emphasis on Assessment *for* Learning in the revision of the assessment regulations was propelled by policy recommendations from the OECD.

Articles II and III illuminate how the role of the OECD was also significant in policymaking in Sweden, related to both grading and testing policies. Swedish policymakers' explicit use of the OECD occurred later than in Norway. As reported in Article III, the government asked the OECD for advice in response to its comprehensive critique of the education system. Its ongoing revision of the national testing programme is partly in response to these inputs. In addition, as reported in Article II, the government nominated an expert to review grading policies in European and OECD countries, who alleged that the countries that were performing better than Sweden in PISA had a system of "early formal grading". The

ministry memorandum referred to an OECD publication that turned out not to exist; instead, the source was primarily information from the Eurydice database. This is an example of how supranational agency sources are used for policy legitimation in tandem with the consultancy mode of policy legitimation, as described below.

The consultancy mode of policy legitimation is less present across sub-studies I, III and IV, but it was highly visible in the reform of grading age in Sweden (researched in sub-study II). The government's legitimation of this assessment reform motivated this thesis' conceptualisation of new modes of policy legitimation in education. Drawing on literature from the United Kingdom, Article II points out that while in some countries (e.g., the United Kingdom) global private enterprises (e.g. McKinsey & Company) have become increasingly involved in policymaking in recent years, it has become more common in Sweden to assign committees or one-man inquiries with a specific mandate to review and recommend policies. A key difference between the role of committees in the collaboracy and consultancy modes is that the latter are based on the government's nomination of experts with fixed mandates that reflect the government's point of view, whereas the views of the former are developed through the democratic participation of the relevant professions.

Furthermore, there is a difference related to the *pace* of these modes of policy legitimation. As observed for Sweden in Article II, there is a remarkable difference between the 1970s, when a committee reviewed and discussed grading age for nearly a decade (before presenting the conclusions and policy recommendations to the government), and the recent one-man inquiry mandated by the Swedish government to review the policies in other countries in less than a year.

However, perhaps the most remarkable finding in this Swedish case of policy legitimation relates to the *construction*—rather than use—of the international arguments. The government attempted to construct "world situations" to legitimise its policies on grading age and abandon the Scandinavian tradition of prohibiting formal grading in primary education. This was legitimised through the nomination of a neuroscience professor who consulted for the government on the proposed policy changes by referring to OECD reports and EU policy descriptions of European countries' grading policies. But these policy legitimation processes did not reflect a genuine respons to established international policy discourses; rather, they provide an example of how nation states may find it convenient to *construct* "world situations" to legitimise reforms.

Based on the four sub-studies, several overall observations can be made related to the second research question: What tensions related to the purposes of assessment can be identified in nation states' legitimation of educational assessment and testing policies?

The identification of the three transnational trends of educational assessment (Table 9) and the associated three roles of educational assessment (Tables 4) helps reveal how the need to legitimate the national testing programmes has contributed to the accumulation of multiple roles for educational assessment in response to different transnational trends at the time of implementation, revision and expansion. Article IV demonstrates that, influenced by the meritocracy, accountability and Assessment *for* Learning trends, the national assessment instruments have accumulated the roles of certifying, governing and supporting learning and instruction. As such, by encompassing all the roles of educational assessment, these transnational trends have contributed to the tensions in nation states' contemporary national testing instruments that are identified in Article III.

Articles I and III shed light on how the Norwegian government shifted the subsidiary role of the tests from the certifying role to the supporting role by emphasising their potential use in formative assessment. As such, it reveals how the government's emphasis on international research and policy discourses had implications for the roles of assessment that were prioritised in the national testing programme. Following the 2005 parliamentary election, the new government wanted the tests to be "suitable as a pedagogical tool" (Article III, p. 336). This requires that the tests are available to teachers and students after they have been taken, as a basis for follow-up, which in turn implies that items cannot be reused, thus imperilling the monitoring of student cohorts' attainment over time. It took ten years before this dilemma was settled by the inclusion of anchor items in 2014. This delay may reflect how the strengthened emphasis on the tests' supporting role from 2006 onwards overshadowed their governing role.

Article IV demonstrates how Sweden implemented standardised national tests as part of the meritocracy trend, and later, as part of the accountability trend, strengthened the governing role of the same testing programme. By contrast, Norway implemented new standardised tests in addition to the existing examinations. Norway was thus spared the tensions caused by integrating the certifying and governing roles in the same instrument. The differences between the contemporary national testing programmes in Norway and Sweden show how the timing of implementation may be one explanation for the two countries' differing emphases on subject-based versus skills-based competences in the national assessment instruments used to govern the education systems.

The combination of all three roles of educational assessment in one instrument in Sweden has caused conflicts with respect to which role should be prioritised. Article III sheds light on ongoing policymaking processes aiming to restrict the purpose of national tests to supporting teachers in their responsibility for grading student attainment (their certifying role).

Article II does not focus directly on the tensions of educational assessment; however, this remarkable Swedish case of policy legitimation illustrates the political controversies and conflicts over what educational assessment should be used for. It is an example of political disputes between the government and the teacher profession, who feel that their supporting role is being undermined by the emphasis on formal grading associated with the use of grades for certifying and governing roles. Article I also points to such conflicts in Norway, illuminating how the tensions between the roles of educational assessment shaped the assessment regulations. Furthermore, this article discusses the Scandinavian legacy of prohibiting formal grading, which can be understood in the light of the tensions between the certifying and governing roles of educational assessment emphasised in educational administration and the teachers' role in supporting students' learning.

In the next chapter, these findings are discussed in relation to the theoretical perspectives and empirical findings of other empirical studies.

6 Discussion

This thesis stimulates key debates about policy borrowing and lending in the legitimation of educational assessment reforms. In this chapter, I first consider the findings related to the first research question regarding how policymakers engage in international research and policy discourses in the legitimation of nation states' educational assessment and testing policies. I discuss the extent to which the OECD's significant role in Scandinavian policymaking and policymakers' constructions of "world situations", which are portrayed as *new modes of policy legitimation* in this thesis, can be perceived as a *Scandinavian governance turn*.

In the second section, I discuss the findings in relation to the second research question regarding tensions related to the purposes of assessment that can be identified in nation states' legitimation of educational assessment and testing policies. I address the conflicting purposes of educational assessment. I then discuss the implications of the epistemological differences related to subject- and skills-based assessments and the disputes over formal grading in Scandinavian education.

The discussion of the study's findings in response to the two main research questions stimulates further discussion of key aspects related to *borrowing and lending in the formative assessment policy and research discourses*. I therefore proceed to discuss definition problems related to the formative and summative assessment distinction, pointing to the implications of the 1980s and 1990s British lending context for the policy borrowing to other contexts and showing how the reciprocal borrowing and lending semantics of educational reforms have diluted the meaning of formative assessment. I discuss the diluted meaning of the roles of evaluation and assessment as a result of the dichotomous use of Scriven's (1967) distinction, as well as its implications for nation states' policymaking.

I conclude by highlighting parallels between the disputes over the use of educational evaluation in the United States in the 1960s and contemporary disputes over educational assessment in Scandinavia. I illuminate that Scriven (1967) aimed to highlight that all evaluation processes are about making judgments of quality or attainment level, irrespective of what this information is used for. By distinguishing between merely a formative and a summative role, however, Scriven undermined this key message about the multiple roles associated with the evaluation processes. I conclude the discussion addressing the implications of this confusion for the legitimation of contemporary policies of educational assessment.

6.1 New modes of policy legitimation

This thesis sheds light on developments in two Scandinavian nation states' legitimation of assessment policies and reforms over the past two decades. Furthermore, it places these developments in a historical context, highlighting the long-standing Scandinavian tradition of collaboration between stakeholders, experts and the teacher profession in education reforms. The thesis also points to developments in policymaking, which are marked by an increased emphasis on the OECD and other supranational agencies, as well as the use of individual consultants, in the legitimation of educational assessment policies. Do these changes reflect a *Scandinavian governance turn*?

6.1.1 The role of the OECD in nation states' policymaking

The notion of educational governance relates in part to the role of supranational agencies. One of the most striking findings of this thesis concerns the significant role of the OECD. Norway has a long tradition of basing its policies on OECD recommendations. Other studies have observed that the new curriculum reform based its definition of competence aims on the OECD's DeSeCo framework (Mausethagen, 2013c). Research on higher education policies have established the large emphasis on policies of another supranational policy actor, the EU. Elken (2016) observed that Norway adopted the European Qualification Framework (EQF) from 2005 to 2009 with a broad consensus among the stakeholders. Dahl and Lindberg-Sand (2009) discussed how one of the implementations of the European Credit Transfer System (ECTS) was manifested in a revision of the higher education grading scale by simply adhering to the ECTS scale.

Kamens (2015) discussed how the OECD's first PISA study, undertaken in 2000, radically changed the premises of policy legitimation and education governance across the globe, causing a "manic search for best practices" (p. 137). With respect to the OECD's role in Sweden, the OECD report of 2015, entitled *Improving schools in Sweden: An OECD perspective* (OECD, 2015), marks a striking contrast to the practices of "silent borrowing" observed by Waldow (2009) in the reforms of the 1960s and 1970s, which favoured legitimation strategies other than explicit borrowing. Ringarp and Waldow (2016) observed that a shift towards utilising international arguments occurred in 2007, which was related to declining results on the PISA tests and the changing perceptions regarding these results in the public discourse. In the area of national testing, but also in the wider education policy discourse, the emphasis on the OECD report of 2015 reflected a breakdown of the "powerful myth of national

superiority" (Waldow, 2009, p. 477), undermining Sweden's "self-confidence as a pioneer country" (Ringarp & Waldow, 2016, p. 6).

Researching five separate OECD reviews of evaluation and assessment practices, Pettersson, Prøitz and Forsberg (2017) demonstrated how "national vertical and/or horizontal developments are intertwined with the OECD policy recommendations" (p. 721). They portrayed this development as *infrastructural governance,* which has established international networks and systems to collect and compare statistical data in education, thereby producing an "epistemological governance, which reflects its well-established capacity to shape the views of key actors in education across local, national and global scales" (Pettersson, Prøitz, & Forsberg, 2017, p. 722). Nordin and Sundberg (2014) discussed how the UNESCO, the World Bank, the OECD and the EU "have come to play an increasingly important role in the construction of transnational policy arenas, as resourceful actors working together, forming powerful discourse coalitions that influence and to some extent even govern national reforms" (p. 14).

Despite, or perhaps because of, this close collaboration, the role of the OECD in Norwegian policymaking faces criticism. One of the most vibrant critical voices in this public discourse is professor emeritus Svein Sjøberg. Making the argument that the "OECD governs Norwegian education", Sjøberg (2014) quoted the former Norwegian prime minister Jens Stoltenberg's speech at the OECD's 50-year anniversary in 2011: "Many countries listen to the main messages of the OECD; thus one can say that the organisation has contributed to changing the world" (Stoltenberg to The Norwegian News Agency, May 25, 2011; referred by Sjøberg, 2014, p. 37, my translation). Sjøberg concluded that "it cannot be said clearer. The OECD provides the premises; the OECD changes the world" (Sjøberg, 2014, p. 37, my translation).

While this thesis illuminates the OECD's significant role in the two nation states' policymaking, it also demonstrates how the governments control the policy processes. The cyclic features of policy borrowing and lending imply that nation states are in the position of setting the OECD's agenda. Researching the 1988, 2002 and 2011 OECD reports with policy recommendations for Norway, Prøitz (2015a) concluded that these recommendations "formed a platform for changes in Norwegian education policy" that "reinforced a results-oriented policy by introducing learning outcomes and assessments designed to improve the learning outcomes of all students and to hold actors accountable" (p. 75). However, further researching the presence of references to the Nordic countries in the OECD's (2013) policy review *Synergies for better learning*, Prøitz (2015a) observed that the Nordic countries were more present than one would have expected given the composition of countries involved in the

review. Prøitz observed that the OECD report gave emphasis to Nordic examples of inclusion and teacher and student involvement that may have contributed to the overall holistic and formative approach to assessment and evaluation promoted by this report. Prøitz concluded that this report "implies a change in focus, which possibly downplays the focus on testing and assessment for accountability" (p. 78). Prøitz (2015a) argued that Norway and the other Nordic countries may have had an impact on OECD recommendations, "promoting holistic approaches that emphasis the need to consider the broader range of factors that influence students' learning and results" (p. 78).

Drawing on Börzel and Panke (2013), Prøitz (2015a) demonstrated a sequential approach of uploading and downloading that shaped the OECD's (2013) policy review. This points to a crucial aspect of nation states' policy legitimation, illuminating how nation states' commitment to working with supranational policy organisations such as the OECD goes beyond the classical policy borrowing mechanism associated with mimic institutional isomorphism processes (DiMaggio & Powell, 1983). This commitment also involves taking an active role in defining the agenda of supranational agencies. The uploading or lending context, in this case Norway, may then later in the process benefit from policy recommendations that are based on standards more aligned with its own polices, as Prøitz indicated was the case for the OECD 2013 policy review. Prøitz (2015a) emphasised the need to go beyond a one-dimensional national perspective to understand "the actors, drivers, initiatives and motives involved in change" (p. 79), pointing to Steiner-Khamsi's (2014) observation that "globalization is not an external force, but rather a domestically induced rhetoric that is mobilized at particular moments of protracted policy conflict to generate reform pressure and build policy coalitions" (p. 157).

These findings suggest that Sjøberg's critique fails to recognise nation states' constitutive power. In particular, resourceful Scandinavian nation states are in the position of setting the agenda for the OECD. Sjøberg (2014) asked "how is the OECD using the PISA project?" (p. 37, my translation). A more nuanced and reasonable question may be: *How are the nation states using the OECD and PISA?*

Sjøberg (2014) expressed a concern that "the PISA project defines the premises of education policy in Norway as well as globally" (p. 41, my translation). This thesis suggests that we should be more concerned about other nations' education policymaking being "hijacked" by the OECD. The modes of policy legitimation proposed in this thesis (Table 8) point to how, in the agency mode of policy legitimation, the effects of mimic isomorphism (DiMaggio & Powell, 1983) may be amplified by the coercive power of supranational agencies

through the conditional benefits that they offer. Robertson (2005) illuminated how the World Bank requires education systems to transform to meet the demands of the global knowledge economy and Western neoliberal fiscal policies (Jones, 2004). This type of agency mode of policy legitimation may effectively force countries to accept supranational agencies' testing and accountability policies in order to receive benefits from, for example, the World Bank (Benveniste, 2002).

Addey (2017) observed how the OECD in recent years has developed its PISA tests to target developing countries, thereby strengthening its "infrastructural and epistemological global governance" (p. 311). I am much more concerned with the implications of the OECD for the independence and cultural integrity of nation states other than the Scandinavian welfare states, which are net contributors to the supranational agencies rather than financially dependent upon them. Nevertheless, the legitimation strategies illuminated in the case of the formal grading age reform in Sweden indicate new (mis)uses of supranational policy information. But in this case too, it was the nation state's use of supranational agency data that was problematic, not the Eurydice platform itself. This is discussed further below.

6.1.2 Constructions of "world situations"

The Eurydice database is an example of how a supranational actor such as the EU facilitates transnational policy flows that "are not reducible to, or explicable in terms of, the intentions and interests of individual member states" (Dale, 2005, p. 125). Rather, it can be seen as an example of infrastructure set up by a supranational agency, whereby policymakers report on their own educational policies and reforms and seek equivalent information from other countries to inform their own policymaking. Through this uploading and downloading (Prøitz, 2015a), a complex "network of reciprocal references" (Schriewer, 1999, p. 23) emerges. This practice can be understood as a *normative isomorphism mechanism* (DiMaggio & Powell, 1983), as the governments' policymakers are attentive to "the conditions and methods of their work" (p. 152), which includes paying attention to one another's ways of describing policies.

In Sweden, following the declining PISA outcomes, the public critique of education created a new situation where externalising to "world situations" became a more attractive way of legitimising reforms (Ringarp & Waldow, 2016). The use of Eurydice data is an example of how transnational and supranational sources can be used to "construct world situations in the form of statistical documentations that are recognised as scientific" (Schriewer, 1988, p. 62). Mediated through the recognised Swedish Official Report (SOU) institution and format, such policy reviews and recommendations can provide scientific legitimacy in line with the

government's ideology. As such, advantage is taken of one of the key features of the traditional collaboracy approach to policy legitimation, the SOU, and use it as a medium "for 'levelling up' values and ideologies—through 'world situations'—to scientific evidence that in turn can inform and legitimate reforms" (Article II, p. 644). Do the new fast modes of policy legitimation observed in Norway and Sweden in this thesis echo developments that have been portrayed in English-speaking countries as a *governance turn*?

6.1.3 New modes of policy legitimation: A Scandinavian governance turn?

Researching the use of ILSA data, Lindblad, Pettersson and Popkewitz (2015) observed that global private enterprises such as McKinsey and Company have become increasingly involved in policymaking in recent years. Such private enterprises utilise ILSA data as a basis for reviewing and recommending policies in their consultancy work for constituencies. Gunter, Hall and Mills (2014) described a trend in the UK that they termed consultocracy, in which "non-elected consultants are replacing political debate conducted by publicly accountable politicians" (p. 519). Using the term *consultancy,* this thesis has illuminated legitimation strategies used in Sweden that are similar to those described in the UK. Rather than commissioning companies to undertake policy reviews, the task has been assigned to a single expert perceived to share the government's ideological position.

Researching governance features of policymaking in England, Sweden, Finland and Scotland, Grek et al. (2009) observed that although interaction with the EU has been growing, the narrative of already being "very good at assuring the quality of its education" was strong among Swedish ministry and executive agency interviewees (p. 13). They observed that Swedish policymakers alleged that Sweden "does not import, borrow or copy ideas or models from anywhere" (p. 13). Nevertheless, in correspondence with countries such as England and Scotland, Sweden has adopted increased market mechanisms, including examples of privatisation and new public management. Furthermore, initiatives from global actors such as the EU and the OECD have led to national discussions as to what is required by a nation and its inhabitants to excel in international competition. The so-called "knowledge economy" legitimises external involvement in national education systems (Forsberg, 2014).

While the legitimation of the grading age reform in Sweden explored in this thesis may be an extraordinary example, it indicates a practice of "bypassing" the profession to legitimate a reform. This is uncharacteristic of the Scandinavian tradition of policymaking, which is "characterized by a tripartite system involving the trade unions, employers' representatives, and the state, with a focus on consensus building" (Elken, 2016, p. 636). Further investigations are

needed to illuminate how governance features of policymaking play out in the Scandinavian nation states, but from the data investigated in this thesis, it appears that Sweden is taking a sharper governance turn than Norway.

6.2 The legitimation of assessment purposes

This thesis has illuminated fundamental differences with respect to how the three roles of educational assessment are distributed across the national assessment instruments in Norway and Sweden. It has identified multiple tensions related to the conflicting purposes of educational assessment.

6.2.1 Conflicting purposes of educational assessment

Erickson (2017) pointed to the recommendations put forward in SOU 2016: 25, where it is proposed that the national tests should have "one aim only, namely, to enhance fairness and equity at the individual level" (p. 140). As such, the committee proposed *certifying* learning and instruction as the main role of the national tests (i.e. the tests should assist teachers in the grading of student attainment). It is proposed that the role of *governing* learning and instruction should be mainly served through a second type of instrument that is sample-based. The role of *supporting* learning and instruction is targeted through a third set of instruments that are labelled "national assessment support" ("nationella bedömningsstöd"; SOU2016: 25; Gustafsson & Ericksson, 2018). These policy changes can be perceived as confirming the substantial problems caused by attempting to integrate the multiple roles of educational assessment in one assessment instrument, as observed in this thesis.

A major issue with the integration of the certifying and governing roles in one instrument is that the instrument used to hold teachers accountable for outcomes is the same instrument that teachers use to determine student attainment. This implies that teachers may be inclined to award higher grades to their students on the tests, as the students' grades are used to determine the quality of their own teaching. As control of teachers' grading practices is strengthened, however, we may see an opposite effect: In fear of being accused of being too lenient in their grading, teachers may feel obliged to impose a stricter approach to their grading of national tests.

Either way, this accountability aspect related to meritocratic procedures may undermine a genuine focus on how to arrive at the correct level of attainment. As such, the combination of governing and certifying roles may be perceived as detrimental both to the process of grading

(certifying role) and to the system-level use of the outcomes of the grading (governing role). The integration of both roles in one instrument results in the two roles undermining each other.

Article III illuminates how Norway's integration of the governing and supporting roles also came with a high price. It implies that test results should be made available to teachers and students so that teachers can use the information about the types of tasks in which the students succeeded as the basis for follow-up. Thus, items cannot be reused, imperilling the monitoring of student attainment over time, which was originally a key objective of the testing programme. It took more than 10 years before this dilemma was resolved through the inclusion of anchor items. This delay can be explained by the strengthened emphasis on the supporting role and the use of the international formative assessment policy and research discourses as arguments for this prioritisation. Furthermore, the purposes of the national testing programmes can be related to the time or era in which they were implemented, as discussed below.

6.2.2 Implications of the timing of embracing standardised testing

Lundahl and Waldow (2009) distinguished between a first and second cycle of standardisation that can be related to the transnational trends of educational assessment discussed in this thesis. In what they called *the first cycle of standardisation*, in the 1930s, standardised tests were introduced in Sweden to individualise instruction and strengthen teacher professionalism. However, their form and function changed as the assessment data moved from the level of instruction to the level of administration. "Instead of being a tool in everyday classroom instruction, they were transformed into bureaucratic and political tools for the establishment of an effective and equal comprehensive school system" (Lundahl & Waldow, 2009, p. 367). In what Lundahl and Waldow (2009) called *the second cycle of standardisation*, in the 1990s, standardised tests served as a means for controlling the decentralisation of the education system. While the first cycle served as a means of *bringing* actors and school organisations together, the second cycle served as a means of *holding* them together.

The identification of transnational trends in this thesis complements Lundahl and Waldow's (2009) perspectives on the first and second cycles of standardisation by highlighting the emphasis on optimising meritocratic procedures in the international research collaborations of the 1930s onwards and the shift towards a greater emphasis on accountability in the international research collaborations of the 1990s onwards. I demonstrate that the implementation and revision of the national testing programmes from the 1930s onwards and in the 1990s can be understood in view of the international research collaborations related to the IEI and IEA studies, respectively.

Moving further ahead in time, to the turn of the millennium, the OECD PISA tests fundamentally changed the premises of international research and policy collaboration and, as such, became the dominant type of supranational tests that characterise the accountability trend. Whereas the supranational IEA tests and the national tests in Sweden were subject-based, the OECD's tests were centred around *skills*, motivated by the agency's emphasis on 21st-century skills. During the first cycle of standardisation and the meritocracy trend, Norway had resisted the supranational IEA tests and rejected the idea of implementing national tests. At the turn of the millennium, however, Norway reacted to the accountability trend by adopting the OECD PISA tests and subsequently implementing national tests of basic skills modelled on the PISA tests. But given that they were based on skills rather than subjects, these tests supplemented rather than replaced the existing examinations.

This suggests that the assessment data used for governing the education system in Sweden is largely based on subject content, whereas in Norway it is more in line with the OECD's skills orientation. It is beyond the scope of this thesis to research the implications of these differences, but it is plausible to ask whether Norway's improved outcomes on the PISA tests, and the corresponding decline in Sweden's outcomes, can be related to these different premises. In other words, in terms of PISA test outcomes, is Norway benefitting from aligning its national testing infrastructure to the OECD PISA tests?

6.2.3 Epistemological differences of subject- and skills-based assessments

Another issue which this thesis sheds light on is how successful the OECD has been in setting the agenda through the PISA tests. Here too, the development of the IEA tests offers an interesting comparison. Nation states' teacher professions maintain control over the *subjects* in ways that they do not with respect to *skills*. Subjects are part of the infrastructure of the education system. They structure the curricula, the teacher education and the certification of teachers. Skills, on the other hand, are not controlled by the teacher profession and are instead a universal and overlapping feature of education systems. The implications of these differences should not be underestimated. Epistemologically, subject knowledge stands in a social constructivist tradition that recognises the complexity and difficulties of agreeing upon attainment. Extensive efforts and resources are put into developing sound tests or examinations that support teachers' capacity to make comparable judgments. Skills, on the other hand, are epistemologically more related to the positivist notion of the capacity of a test to determine a "true score". Figure 1 coins a distinction between these two epistemological paradigms: *professional social judgment* versus *external objective measurement*.

With respect to the borrowing and lending features of nation states' policymaking discussed in this thesis, it can be argued that skills-based tests, operating in the so-called external objective measurement paradigm, are easier for nation states to agree upon as they are not rooted in subjects that are controlled by the teacher profession. Furthermore, the notion of a "true score", or at least the idea that the scores are truer than teacher judgments, implies that such tests have higher scientific credibility. This scientific aura can further be associated with the ILSA research collaboration, which has established a common scientific language about validity and comparability of assessments. This is an example of normative isomorphic mechanisms identified by DiMaggio and Powell (1983), whereby scholars "define the conditions and methods of their work" (p. 152). As a result, these types of tests may be perceived as more legitimate measures of attainment than teacher judgments, whicht are limited to national professional language. Consequently, testing instruments closer linked to supranational instruments are perceived more appropriate for governing education systems.

6.2.4 Disputes over formal grading in Scandinavian education

This thesis has demonstrated that policies for educational assessment need to tackle fundamental issues related to the problems of validity and comparability of judgments (certifying) and the system-level use of assessment information (governing). I contend that this premise is crucial to understand the Scandinavian legacy of avoiding this legitimacy threat by simply prohibiting formal grading. The formative assessment discourse, which points to the negative effects of grading and "summative assessment", thus fits well with the ideological backbone of Scandinavian education.

Current assessment policy tensions in Norway and Sweden draw on similar ideological arguments that once shaped the Scandinavian tradition of prohibiting mark-based assessments in primary education. The rationales for formal grading typically follow three lines of argument: motivation, selection and information (Wikström, 2006). In Sweden, formal grading was prohibited in primary school in 1962, while in Norwegian primary schools, it has been prohibited since 1973. Concerns about the negative impact of formal marking on low achievers have since been the main reason for continuing the prohibition of grading in the primary sector (Tønnessen & Telhaug, 1996). The most prevalent rationale for grading students is for purposes of selection (certifying learning and instruction). However, as students go to designated schools in their districts throughout the compulsory years of education, no selection procedures are necessary until they transfer from lower to upper secondary education. In the 1970s, there were strong advocates of prohibiting formal grading in secondary education as well; however, these

attempts failed to offer plausible alternatives for selecting students for further education (Lysne, 2006).

The disputes over formal marking in the 1970s did, however, have far-reaching implications for determining attainment in primary education and for grading practices in secondary education. They substantiated scepticism towards explicitly stated learning objectives that could be measured and controlled, and both left- and right-wing governments complied with this when revising the curricula in the 1980s and 1990s. The third argument, that formal grades provide helpful information to students and parents as to where students are in their learning, may thus have had less relevance in the Scandinavian education systems, as grades have not traditionally been determined based on clearly stated learning outcomes. This has, however, changed during the past two decades. The standards-based curriculum in Sweden of 2011 states levels of attainment, while in Norway the curriculum of 2006 clearly articulates the expected learning outcomes as competence aims. Thus, both in Norway and Sweden, we can observe more emphasis on facilitating formal grading, in concert with the increased emphasis on formative assessment. The increased emphasis on both the certifying and supporting roles may seem as a paradox. The next section further addresses this link by discussing how borrowings from the formative assessment policy and research discourses can be related to the Scandinavian scepticism towards formal grading.

6.3 Borrowing and lending in the formative assessment policy and research discourses

I conclude this discussion chapter by discussing *borrowing and lending in the formative assessment policy and research discourses*, making the case that the influence of these discourses reflects strategies for legitimating wider educational assessment reforms related to both the certifying and governing roles of educational assessment. In conventional terms, that is, I discuss how formative assessment discourses de facto are used to legitimise summative assessment.

This thesis has demonstrated that policies for educational assessment need to tackle fundamental problems for which there is no ultimate solution. Problems of validity and comparability of judgments, together with the use of assessment to control the teacher profession, imply that the legitimacy of educational assessment policies is constantly threatened. This, in turn, may explain why policy borrowing is widely used in nation states' legitimation of educational assessment policies. That is, policies that entail control over the

teacher profession and increased workloads for teachers become more "popular" when implemented as formative assessment principles or as part of AfL programmes. Due to the popularity of formative assessment, its theoretical interpretations and practical applications, and the embedded definition problems, are widespread. These borrowing and lending problems may ultimately backfire and undermine the legitimacy of these assessment policies.

6.3.1 Definition problems of the formative and summative distinction

Based on a close reading of the definitions of formative assessment, Wiliam (2011) acknowledged that "there is no clear consensus about the meanings of the term" (p. 9). This is remarkable given Black and Wiliam's optimistic reporting on the potential implications of formative policies and practices (1998b) and their efforts to develop a theory of formative assessment (2009; Chapter 3.2.3). Chapter 2.9 reviewed researchers use of the distinction between formative and summative assessment, identifying five different ways of defining the relationship. Whereas the early formulation of the distinction (Bloom et al., 1971; Sadler, 1989) made a basic distinction related to timing, other researchers (Black, 1998; OECD, 2005; Stobart, 2008) distinguished between the use of summative assessment for the certification of individual learning and the evaluation of teachers and schools. Black and Wiliam (1998a, 2009), however, do not offer an explicit definition of summative assessment.

Biggs (1998) criticised Black and Wiliam (1998a) for excluding summative assessment from their review of the effects of assessment on classroom learning. Biggs pointed out the implicit mutually exclusive understanding of formative and summative assessment resulting from Black and Wiliam's definition of formative assessment. According to Black and Wiliam (1998a), feedback is only regarded as "formative" when "comparison of actual and reference levels yields information that is then used to alter the gap" (p. 53). Biggs (1998) pointed out that this implies that "if the information cannot lead to appropriate action—it becomes a summative grade, for instance—then it is not formative. . . . They [formative and summative assessment] are seen in effect as mutually exclusive" (p. 106, my clarification in brackets).

Reviewing the use of the distinction between formative and summative assessment in the assessment literature, Lau (2016) observed that the literature has increasingly condemned summative assessment. "A number of models, paradigms and conceptual frameworks are being put forward in an attempt to engender a move away from summative assessment" (p. 513). The EARLI position paper referred to in chapter 2.8 and Table 1 (Birenbaum et al., 2006) is illustrative of the dichotomous use of the formative and summative assessment distinction—and the associated emphasis on Assessment *for* Learning versus Assessment *of* Learning. This

mutually exclusive (either this or that) rhetoric distorts the practical use of the concepts and undermines comparative analyses of assessment policies.

Taras's (2007) interpretation of Sadler (1989) is that the concept of a goal (or *standard* in Sadler's terminology) is intrinsic to formative assessment and that summative and formative assessment logically lead into each other as one continuous process. While I recognise Taras's critique of Black and Wiliam's (1998a, 2009) failure to define summative assessment, I view the solution put forward by Taras as just another perspective on how the formative and summative assessment relationship *should* be perceived. As addressed in the review chapter (Sections 2.7 and 2.8), the theoretical definitions of formative assessment and nation states' implementation of AfL programmes and other formative assessment policies have been criticised in recent years. Without discrediting research studies and practices labelled as formative assessment or AfL, it is fair to say that the vast emphasis on the concept in policy and research has accumulated to a definition problem.

The ambiguous meaning of formative assessment is an example of how "reciprocal references" emerge from the accumulation of observations of various nations, acquiring their own autonomy that "transmits, confirms and accelerates the planetary universalising of reform representations" (Schriewer, 1999, pp. 23–24.). The prominent lending context in the new millennium has been the United Kingdom and the work of Black and Wiliam (1998a, 1998b). The next sections discuss the practices of policy lending and policy borrowing in the United Kingdom and Norway, respectively, examining how this has contributed to the "transnational semantics of pedagogical reform" (Schriewer, 1999, p. 24).

6.3.2 The British renaissance of formative assessment discourse

Black and Wiliam (2003) portrayed how their efforts to promote renewed emphasis on formative assessment in policy and research were motivated by the perceived extreme focus on tests that accompanied the introduction of the national curriculum in England and Wales in 1988. Excessive testing caused substantial concern for British educators and researchers at the time (Isaacs, 2010).

The British Educational Research Association (BERA) established an Assessment Policy Task Group in the 1990s as a direct attempt to "rework the policy discourse" (Daugerthy, 2007, p. 145). The Assessment Reform Group (ARG) was worried that the meaning of formative assessment would be hijacked to legitimise teacher accountability policies. Similar problems related to the popularisation of the concept in both policy and commercial areas can

be observed in the United States.[8] The ARG was concerned that some practices were perceived to be formative simply because the assessments were constructed and administered by teachers rather than external agencies. The AfL principles were designed to express a contrast with

> assessment that simply adds procedures or tests to existing work and is separated from teaching, or on-going assessment that involves only marking and feeding back grades or marks to pupils. Even though carried out wholly by teachers such assessment has increasingly been used to sum up learning, that is, it has a summative rather than a formative purpose. (ARG, 1999, p. 7)

Funded by the Nuffield Foundation, BERA commissioned Black and Wiliam to undertake a review of the research literature, "convinced that research evidence on the effectiveness of formative assessment practices would demonstrate their potential both for deeper learning and higher standards of attainment" (Daugerthy, 2007, p. 145). Black and Wiliam's (1998a, 1998b) publications received tremendous global attention. Kirton, Hallam, Peffers, Robertson and Stobart (2007) noted that the pamphlet *Inside the Black Box* (Black & Wiliam, 1998b), which summarised the key messages of the review, sold 50,000 copies in its first ten years. Overall, the reviewed research studies brought an increased attention to how formal assessments (tests and examinations) can be detrimental to learning. This sparked an increased focus on the learning aspects of assessment, often described as formative assessment or AfL, which has since become important in many governments' legitimation of educational assessment policies.

Black and Wiliam (2003) recognised that a main motivation of the 1998 review was to confront the development whereby teachers were left to administer tests and to (re-)establish assessment as a classroom activity centred on learning. In a more recent publication, Black (2015) problematized how the understanding of formative assessment and AfL policies emerged following his joint publication with Wiliam in 1998:

> Many writers about assessment, and many teachers, regard assessment as a peripheral component of pedagogy, one that is inescapable but which always threatens to undermine the most valued aim, that of developing the learning capacity of their students. The phrase "assessment for learning" challenges this view, and some handle this challenge by regarding it as quite separate from summative assessment. This I regard as a fundamental error, one that arises from the lack of a broad and more complex view of the role of assessment in pedagogy. (Black, 2015, p. 163)

Lau (2016) argued that a "formative good, summative bad" dichotomy was "unintentionally created by those promoting assessment for learning" (p. 512). This was largely related to the

[8] Shephard (2006) observed that "the research-based concept of formative assessment, closely grounded in classroom instructional processes, has been taken over (hijacked) by commercial test publishers and used instead to refer to formal testing systems called benchmark or interim assessment systems" (referred by Popham, 2011, p. 275).

political and practical context in the UK. When research and policy discourses situated in the UK setting became a defining feature of the transnational policy and research discourses, including reviews and recommendations by the OECD, these concepts fed into contexts vastly different from the lending context. This is problematic given the reciprocal borrowing and lending semantics of nation states' policymaking, which dilute the meaning of formative assessment, as experienced in the case of the reform of the assessment regulations in Norway discussed below.

6.3.3 The Norwegian borrowing and lending context

The curriculum reform in Norway in 2006, and the accompanying revision of the assessment regulations in 2009, is a remarkable context of borrowing and lending, as it did not merely include emphasise on formative assessment in policy documents and political rhetoric but also altered the statutory regulations for teachers' everyday work. For decades, the Norwegian Education Act and associated assessment regulations were structured based on two key concepts: "formal assessment" (Norwegian: *formell)* and "informal assessment" (Norwegian: *uformell*). When new regulations were implemented, the formal/informal distinction was discontinued and replaced by a new distinction between *underveisvurdering* ("under-way", "ongoing" or "continuous" assessment) and *sluttvurdering* ("final" or "end" assessment).

The term formative assessment materialised in the 2009 assessment regulations as "underveisvurdering". It is translated as *formative* assessment in the official English dictionary,[9] and it is sometimes supplemented with the clarification "continuous" in brackets (e.g. in the executive agency's reporting to the OECD [DET, 2011] and the EU Eurydice platform).[10] This implies that any assessment that is *not* undertaken towards an end is called formative assessment. This dilution of the meaning of formative assessment continues: In the ongoing 2020 curriculum reform[11], each subject curriculum describes requirements for "underveisvurdering" (formative assessment). Many of the requirements outlined are basically teaching guidelines that perhaps more precisely could be described as "classroom assessment" rather than "formative assessment". In this way, the meaning of formative assessment vanishes.

9 https://www.udir.no/arkivmappe/Ordbok/
10 https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-single-structure-education-20_en
11 https://www.udir.no/laring-og-trivsel/lareplanverket/fagfornyelsen/

This is a problem not limited to the Norwegian context. Moreover, the situation points to the challenges of *lending* reformed policy contexts, using ambiguous representations of concepts such as formative assessment in, for example, OECD and EU policy papers. As discussed in Section 6.1.1, one "driver" of the international policy discourse is the interest of nation states in taking an active role in setting the agenda of supranational agencies such as the OECD. This becomes a problem, however, when the meaning of the borrowing and lending concepts becomes diluted. The use of formative assessment as a legal term implies a vast use of the concept, both within the national context and when reporting in English in policy and research contexts. As such, Norwegian policymakers and researchers contribute to the dilution of the meaning and understanding of formative assessment worldwide.

6.3.4 The diluted meaning of formative assessment

According to Schriewer (1999), a "transnational semantics of pedagogical reform accelerates reform representations and models" (p. 24). As a result, we risk undermining the legitimate practices that formed the basis of the initial lending and borrowing of so-called formative assessment policies, with the message becoming diluted and confused as the concepts are reflected and defined within national education systems. The borrowing and lending of the formative assessment concept between the United Kingdom and Norway illustrates how its meaning has become diluted.

Acknowledging this dilution in the contemporary formative assessment policy and research discourses, it is imperative to look at the origin of the concept. More than 50 years have passed since Scriven (1967) coined the distinction between formative and summative evaluation. While Scriven (1967) is usually referred to as the origin of the distinction, few researchers quote or discuss his definitions or the context of the publication. I will conclude this discussion asking if the diluted meaning and confused message is the result of the transnational borrowing and lending semantics of policymaking and assessment reforms, or whether the conceptual confusion is inherent in the distinction itself.

Scriven's (1967) paper, *The Methodology of Evaluation*, coined the distinction in a four-page section entitled "Goals of evaluation versus roles of evaluation: Formative and summative evaluation". A close reading of these pages unravels forgotten messages in Scriven's paper. Scriven used the term "goals of evaluation" to describe the activity of evaluation (i.e. the activity of determining goal [or standard] attainment), while the term "roles of evaluation" refers to what these evaluations are used for (see Table 4).

Scriven expressed a concern related to the growing opposition towards evaluation as a result of the negative implications of what he called the summative role. This was part of a discussion with professor Lee J. Cronbach, who was concerned about the negative implications of evaluation. Cronbach (1964/1983) argued that "evaluation, used to improve the course while it is still fluid, contributes more to improvement of education than evaluation used to appraise a product already placed on the market" (p. 105). Scriven (1967) was provoked by Cronbach's foregrounding of one usage of evaluation, arguing that "fortunately, we do not have to make this choice" (p. 43). It was as part of this line of reasoning that he distinguished between two roles of evaluation (see Table 4). Scriven was worried that the rhetoric of Cronbach (1964/1983) would make people oppose evaluation per se. He had the following to say regarding making evaluations without any concept of levels or standards:

> By stressing the constructive part evaluation may play in nonthreatening [formative] activities (roles) we slur over the fact that its goals always include the estimation of merit, worth, value, etc., which all too clearly contribute in another [summative] role. . . . But we cannot afford to tackle anxiety about evaluation by ignoring its importance and confusing its presentation. (1967, p. 42, my clarifications in brackets).

Scriven (1967) highlighted that "the *role* which evaluation has in a particular educational context may be enormously various", listing a dozen examples ranging from evaluation of teacher training activities to rewarding or punishing individuals in a prison or a classroom (pp. 42–43). Scriven's point in distinguishing between the goals and roles of evaluation was that *any* evaluation entails the determination of goal attainment, and that it is not reasonable to reject activities for determining goal attainment per se (what I call the assessment process) simply because some uses of evaluation data are unacceptable. Scriven (1967) contended that "failure to make this rather obvious distinction between the roles and goals of evaluation is one of the factors that has led to the dilution of the process of evaluation. . . . This dilution has sacrificed goals to roles" (p. 41).

I will conclude this thesis making a claim that Scriven's key point in recognising the *many roles of evaluation* was undermined by his own distinction between formative and summative evaluation. Ironically, the seed of the mutually exclusive, and thus dichotomous, understanding of the two forms of assessment was sown in Scriven's own formulation of the distinction, which in turn was soon embraced by Bloom (1968), Bloom et al (1971) and other influential American scholars. Looking back at the extensive use of the distinction, Scriven (1990) described its construction as a successful process of neologising, where in particular the term *summative* had an instant "self-explanatory" and comprehensible meaning as a result of its "juxtaposition with formative", even though it could not be found in any dictionary (p. 26). Given its widespread use, it is evident that the distinction came across as highly intuitive.

Rereading Scriven (1967) and his retrospective (1990) account of the distinction, I however contend that the idea of "summative" as self-explanatory is highly problematic. It may seem self-explanatory within *one* education system or assessment culture. However, in contemporary policymaking, shaped by transnational borrowing and lending—and increasing influence from supranational agencies such as the OECD—this is a fundamentally naïve position. "Self-explanatory" approaches are bound to be context-dependent, which implies that the self-explanatory basis of the distinction takes different forms in different contexts. Scriven's neologism for the summative side of the distinction—together with the lack of attention by Bloom (1968), Black & Wiliam (1998a, 1998b, 2009) and others to the theoretical foundations of the distinction when popularising it in the area of educational assessment—has caused conceptual problems to which all users of the concepts have since become victim.

It is a paradox that perhaps the most substantial contributor to sacrificing the goals of evaluation to its roles was Scriven himself, through his coining of this distinction between just two roles. This caught everyone's attention at the expense of his real distinction and crucial point that *judgments of quality are part of* any *evaluation process irrespective of what the information about the determined quality or attainment level is used for.* It is my hope that this thesis, more than 50 years later, will clarify Scriven's key message in relation to the assessment of individual students: *The determination of goal or standard attainment is characteristic of* any *assessment process, irrespective of what the information about the determined goal or standard attainment is used for.*

7 Implications of the Study

This thesis contributes to the policy borrowing and lending field by demonstrating both theoretically and empirically that educational assessment is particularly vulnerable to this phenomenon. Given that nation states and their policymakers, teacher professions and other stakeholders are de facto unlikely to be able to agree on how to carry out assessments, a legitimacy crisis exists that propels the use of externalisation to world situations as a reform strategy. In this concluding chapter, I address seven main implications of the study related to: (1) the theoretical contributions for comparative research; (2) the professional judgments versus external measurement typology; (3) the Scandinavian legacy of prohibiting formal grading; (4) the legitimacy crisis of the quest for just assessments; (5) the idiosyncratic contexts of formative assessment; (6) whether the thesis portrays a blow to the Nordic model; and conclusively (7), exposing problems of the formative and summative assessment distinction.

I conclude by highlighting parallels between the disputes over the use of educational evaluation in the United States in the 1960s and the Scandinavian legacy of disputes over formal assessments that continue up to this day, relating the formative assessment research and policy discourses to these ideological discussions of key principles.

7.1 Theoretical contributions for comparative research

The thesis has promulgated three analytical frameworks that can be used to research and compare assessment policies in other studies and contexts beyond Scandinavia. The distinction between three *roles of educational assessment* (assessments used to c*ertify*, *govern* and *support* learning and instruction) and the associated distinction between *transnational trends* (*meritocracy, accountability* and *Assessment for Learning*) may provide suitable frameworks for contemporary and historical research studies of assessment policies elsewhere. Furthermore, the distinction between *three modes of policy legitimation* (*collaboracy, agency* and *consultancy*) may foster research studies on policymaking, policy legitimation, and policy borrowing and lending in other research disciplines (e.g. political science).

7.2 The *professional judgments* versus *external measurement* typology

The theoretical framework distinguishing between *professional (subjective) judgments* and *external (objective) measurement* (Figure 1) contributes to the perspectives opened up by Hopmann (2003), with *professional judgments* and *external measurement* corresponding to Hopmann's distinction between process- and product-controlled education systems, respectively. Combined with historical perspectives related to transnational trends of educational assessment, this typology may be used to research the epistemological orientation of nation states' assessment policies. Based on the present study, there is a basis for speculating that comparable traits exist between Norway and Germany, given Germany's corresponding resistance to standardised testing and its sustained examination tradition. The Netherlands, on the other hand, can be assumed to share the legacy and contemporary epistemological orientation of Sweden, being another early adopter of standardised testing. These traits should be explored further in research studies on educational assessment policies. Furthermore, perspectives from the sociology of knowledge may foster further research into these suggested epistemological patterns in nation states' educational assessment policies.

7.3 The Scandinavian legacy of prohibiting formal grading

The thesis sheds light on the Scandinavian legacy of prohibiting formal grading in primary education and the genuine overall scepticism towards grading and standardisation. It further shows that there has been an attempt to legitimise ideological arguments by (mis)using comparative data to construct "world situations" with respect to formal grading policies. As such, the thesis has revealed and opened up the ideological disputes associated with formal grading in Scandinavian education policies. The thesis furthermore links this legacy to the renaissance of formative assessment, which was embraced through borrowing from predominantly UK–driven formative assessment policy and research discourses. This is a paradox, given that the values associated with formative assessment have always been core values in Scandinavian education. As such, the thesis explains why the international formative assessment policy and research discourses, and Assessment *for* Learning, have been embraced so wholeheartedly in Scandinavia, as the Norwegian case in particular demonstrates.

7.4 The legitimacy crisis of the quest for just assessments

The thesis illuminates how the legitimacy of educational assessment policies relates largely to the comparability of assessments. Moreover, it establishes that it is virtually impossible to achieve fully comparable and just assessments. The changes that are currently being undertaken with respect to the use of national tests in Sweden reflect an emphasis on fairness that can be related to politicians' commitment to ensuring public (i.e. students' and parents') confidence in the assessments. However, the thesis highlights that this promise can never be met. Therefore, this continues to be a constant threat to the legitimacy of educational assessment policies, as well as a threat to politicians' re-election. This legitimacy crisis may explain politicians' increased emphasis on standardised testing, as it can paint a picture of more scientific and sound assessment processes, as well as the tendency of governments to use externalisation to "world situations" and policy borrowing as legitimation strategies.

7.5 The idiosyncratic contexts of formative assessment

The study has illuminated how the unclear relationship between formative assessment and summative assessment in their main contributions (Black & Wiliam, 1998a, 1998b, 2009) has diluted and confused Black and Wiliam's widespread message regarding formative assessment. The various interpretations of summative assessment, including its "self-explanatory" meaning, undermine a shared understanding of the concepts across contexts. In the example of the United Kingdom as a lending context, a plausible question is whether the perceived excessive summative assessment practices and their negative effects are related to an over-emphasis on testing and examinations (the certifying role) and/or strengthened accountability policies that undermine teacher autonomy (the governing role). Due to a failure to define summative assessment, fundamental differences with respect to the roles of certifying and governing learning and instruction are unclear in the formative assessment policy and research discourses. Thus, policymakers may find themselves borrowing policies that are responding to issues remote from their own contexts of implementation.

The example of the revision of the assessment regulations in Norway illustrates that legal language is shaped by the international research and policy discourses. As such, conceptual tensions between formative and summative assessment in international researchers' and policymakers' use of the distinction feed into the Norwegian political, legal and practical contexts and are in turn uploaded to the international policy and research discourses through

translations and policy descriptions. A result of this transnational semantics of pedagogical reform is that the highly idiosyncratic reform representations are constantly deepened. Thus, while there may be an attractive veneer of resemblance and a shared language for policymakers and researchers across various education contexts, in fact, the premises for this comparability vanish.

7.6 A blow to the Nordic model of education?

This thesis may be seen as nuancing, or perhaps even calling into question, the very idea of the Nordic model of education. The thesis sheds light on fundamental differences between Norway's and Sweden's educational assessment policies that can be related to key epistemological concepts. Furthermore, it has established that the premises for governing the education system in Sweden have changed radically over the past three decades, with the country experiencing a sharper *governance* turn than its western neighbour. Expanding on the research undertaken in this thesis to include the other Nordic countries may reveal that the notion of a Nordic model in education is about to vanish.

7.7 Exposing problems of the formative and summative assessment distinction

Finally, this thesis exposes problems of the vastly used distinction between formative and summative assessment. Identifying five different ways of defining the distinction, it illuminates that consensus cannot be achieved. Furthermore, the thesis highlights problems of using these concepts when researching, and especially when comparing, educational assessment policies. Furthermore, it establishes that the distinction between formative and summative evaluation was coined by Scriven (1967) in relation to disputes over the use of evaluation in education. The thesis highlights an intriguing parallel: The fundamental confusion related to the meaning of formative and summative assessment is rooted in rhetorical arguments and disputes about the *negative implications of evaluation* in the United States in the late 1960s. Thus, it is perhaps not surprising that more than 50 years later these rhetorical concepts are confusing the policy and research discourses in Scandinavia, one of the regions of the world where the rhetorical claims and disputes about the *negative implications of assessment* are particularly pronounced.

8 References

Addey, C. (2017). Golden relics & historical standards: how the OECD is expanding global education governance through PISA for Development. *Critical Studies in Education*, *58*(3), 311–325. https://doi.org/10.1080/17508487.2017.1352006

Andersen, J. A. (2009). *Organisasjonsteori. Fra argument og motargument til kunnskap* [Organisation theory: From argument to knowledge]. Oslo, Norway: Universitetsforlaget.

Andresen, S., Fossum, A., Rogstad, J., & Smestad, B. (2017). På prøve. Evaluering av matematikkeksamen på 10. trinn våren 2017 [Tested. Evaluation of mathematics examination for year 10 spring 2017]. Oslo: FAFO.

ARG (1999). *Beyond the black box.* Nuffield Foundation, United Kingdom: Assessment Reform Group. Retrieved from http://www.nuffieldfoundation.org/sites/default/files/files/beyond_blackbox.pdf

Baird, J.-A., Hopfenbeck, T. N., Newton, P. E., Stobart, G., & Steen-Utheim, A. T. (2014). *Assessment and learning: State of the field review.* Oslo, Norway: Knowledge Centre for Education.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25.

Benveniste, L. (2002). The political structuration of assessment: Negotiating state power and legitimacy. *Comparative Education Review*, *46*(1), 89–118.

Berger, P. L., & Luckmann, T. (1966). *The social construction of reality*. New York, NY: Penguin books. Retrieved from http://perflensburg.se/Berger%20social-construction-of-reality.pdf

Bergeron, P. J., & Rivards, L. (2017). How to engage in pseudoscience with real data: A criticism of John Hattie's arguments in *Visible Learning* from the perspective of a statistician. *McGill Journal of Education, 52*(1).

Biggs, J. (1998). Assessment and classroom learning: A role for summative assessment? *Assessment in Education: Principles, Policy & Practice*, *5*(1), 103–110.

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., … Nickmans, G. (2006). A learning integrated assessment system. *Educational Research Review*, 1(1), 61–67.

Bjørnset, M., Fossum, A., Rogstad, J., Smestad, B., & Talberg, N. (2018). *Digitale skillelinjer: Evaluering av matematikkeksamen på 10. trinn våren 2018* [Digital divide: Evaluation of mathematics examinations for year 10 spring 2018]. Oslo: FAFO.

Black, P. (1998). Testing: Friend or Foe? Theory and practice of assessment and testing. London: Falmer Press

Black, P. (2015). Formative assessment--an optimistic but incomplete vision. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 161–177. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/0969594X.2014.999643

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the Classroom. *Phi Delta Kappan*, *86*(1), 8–21.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London: Granada Learning. Retrieved from https://market.android.com/details?id=book-_hDy-5KOro8C

Black, P., & Wiliam, D. (2003). "In praise of educational research": Formative assessment. *British Educational Research Journal*, *29*(5), 623–637.

Black, P., & Wiliam, D. (2005). Lessons from around the world: How policies, politics and cultures constrain and afford assessment practices. *Curriculum Journal*, *16*(2), 249–261. http://doi.org/10.1080/09585170500136218

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability* (formerly: *Journal of Personnel Evaluation in Education*), *21*(1), 5–31.

Bloom, B. S. (1968). Learning for mastery. Instruction and curriculum. *Evaluation Comment*, *1*(2).

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw-Hill.

Bogner, A., & Menz, W. (2009). The theory-generating expert interview: Epistemological interest, forms of knowledge, interaction. In A. Bogner, B. Littig, & W. Menz (Eds.), *Interviewing experts* (pp. 43–80). London, UK: Palgrave Macmillan.

Börzel, T. A., & Panke, D. (2013). Europeanisation. In M. Cini & N. P. Borragan (Eds.), *European Union politics* (pp. 115–127). Oxford, UK: Oxford University Press.

Bray, M., & Kai, J. (2007). Comparing systems. In M. Bray, B. Adamson, & M. Mason (Eds.), *Comparative education research: Approaches and methods* (pp. 123–144). Dordrecht, the Netherlands: Springer.

Brookhart, S. M. (2015). Graded achievement, tested achievement, and validity. *Educational Assessment*, *20*(4), 268–296. http://doi.org/10.1080/10627197.2015.1093928

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*(2), 105–121.

Bryman, A. (2012). *Social research methods* (4th ed.). Oxford, UK: Oxford University Press.

Carson, J. (2006). The measure of merit: Talents, intelligence, and inequality in the French and American Republics, 1750–1940. Princeton, NJ: Princeton University Press.

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics, and challenges. In G. J. Cizek & H. L. Andrade (Eds.), *Handbook of formative assessment* (pp. 3–17). New York, NY: Routledge.

Cliffordson, C. (2004). Betygsinflation i det målrelaterade betygsystemet [Grade inflation in the goal-referenced grading system]. *Pedagogisk Forskning i Sverige*, *9*(1), 1–14.

Cowen, R., & Kazamias, A. M. (2009). *International handbook of comparative education*. London, UK: Springer.

Cronbach, L. J. (1964/1983). Course Improvement through Evaluation. In G. F. Madaus, M. S. Scriven, & D. L. Stufflebeam (Eds.), *Evaluation Models: Viewpoints on Educational and Human Services Evaluation* (pp. 101–115). https://doi.org/10.1007/978-94-009-6669-7_6

Dahl, B. L., & Lindberg-Sand, Å. (2009). Conformity or confusion? Changing higher education grading scales as a part of the Bologna Process: The cases of Denmark, Norway and Sweden. *Learning and Teaching*, *2*(1), 39–79 doi:10.3167/latiss.2009.020103

Dale, R. (2000). Globalization and education: Demonstrating a "common world educational culture" or locating a "globally structured educational agenda"? *Educational Theory*, *50*(4), 427–448. http://doi.org/10.1111/j.1741-5446.2000.00427.x

Dale, R. (2005). Globalisation, knowledge economy and comparative education. *Comparative Education Review*, *41*(2), 117–149. https://doi.org/10.1080/03050060500150906

Daugherty, R. (2007). Mediating academic research: The Assessment Reform Group experience. *Research Papers in Education*, *22*(2), 139–153.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627–668.

Deci, E. L., & Ryan, R. (1985). Intrinsic motivation and self-determination in human behavior. New York, NY: Plenum.

DET. (2011). OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes. Country Background Report for Norway. Oslo, Norway: Directorate of Education and Training.

DET. (2019). *Kunnskapsgrunnlag for evaluering av eksamensordningen* [Research review for the evaluation of the examination system]. Retrieved from https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Kunnskapsgrunnlag-for-evaluering-av-eksamensordningen/. Oslo, Norway: Directorate of Education and Training.

Dexter, L. A. (1971/2006). *Elite and specialized interviewing*. London: ECPR Press.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, *48*(2), 147–160.

Elken, M. (2016). "EU-on-demand": Developing national qualifications frameworks in a multi-level context. *European Educational Research Journal*, *15*(6), 628–643.

Epstein, E. (1988). The problematic meaning of "comparison" in comparative education. In J. Schriewer & B. Holmes (Eds.), *Theories and methods in comparative education* (pp. 3–23). Frankfurt am Main, Germany: Peter Lang.

Erickson, G. (2017). Experiences with standards and criteria in Sweden. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard setting in education: The Nordic countries in an international perspective* (pp. 123–142). Cham, Switzerland: Springer International Publishing.

Forsberg, E. (2014). Utbildningens bedömningskulturer i granskningens tidevarv [Evaluation cultures of education in the audit era]. *Utbildning och Demokrati*, *23*(3), 53–76.

Forsberg, E., & Román, H. (2014). The art of borrowing in Swedish assessment policies: More than a matter of transnational impact. In A. Nordin & D. Sundberg (Eds.), *Transnational policy flows in European Education: The making and governing of knowledge in the education policy field* (pp. 183–208). Oxford, UK: Symposium Books.

Gamlem, S. M., & Smith, K. (2013). Student perceptions of classroom feedback. *Assessment in Education: Principles, Policy & Practice*, *20*(2), 150–169.

Grek, S. (2013). Expert moves: international comparative testing and the rise of expertocracy. *Journal of Education Policy*, *28*(5), 695–709.

Grek, S., & Lawn, M. (2009). A short history of Europeanizing education: The new political work of calculating the future. *European Education*, *41*(1), 32–54.

Grek, S., Lawn, M., Lingard, B., Ozga, J., Rinne, R., Segerholm, C., & Simola, H. (2009). National policy brokering and the construction of the European Education Space in England, Sweden, Finland and Scotland. *Comparative Education*, *45*(1), 5–21. http://doi.org/10.1080/03050060802661378

Gunter, H. M., Hall, D., & Mills, C. (2014). Consultants, consultancy and consultocracy in education policymaking in England. *Journal of Education Policy, 30*(4), 518–539.

Gustafsson, J.-E., & Erickson, G. (2013). To trust or not to trust? Teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability, 25*(1), 69–87.

Gustafsson, J.-E., & Erickson, G. (2018). Nationella prov i Sverige: tradition, utmaning, förändring [National tests in Sweden: tradition, challenges, change]. *Acta Didactica Norge, 12*(4), 1–20. http://dx.doi.org/10.5617/adno.6434

Hayward, E. L. (2007). Curriculum, pedagogies and assessment in Scotland: The quest for social justice. "Ah kent yir faither." *Assessment in Education: Principles, Policy & Practice*, *14*(2), 251–268.

Hayward, L. (2015). Assessment is learning: The preposition vanishes. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 27–43.

Hall, J. B., & Sivesind, K. (2014). State school inspection policy in Norway and Sweden (2002–2012): A reconfiguration of governing modes? *Journal of Education Policy*, *30*(3), 429–458. http://doi.org/10.1080/02680939.2014.945488

Hattie, J. (2009). *Visible learning*. London, UK: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.

Hirsch, Å., & Lindberg, V. (2015). *Formativ bedömning på 2000-talet: en översikt av svensk och internationell forskning*. Delrapport från skolforsk-projektet [Formative assessment since 2000: an overview of Swedish and international research. Interim report from the skolforsk project]. Stockholm, Sweden: Swedish Research Council.

Hopfenbeck, T. N., Petour, M. T. F., & Tolo, A. (2015). Balancing tensions in educational policy reforms: Large-scale implementation of Assessment for Learning in Norway. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 44–60.

Hopfenbeck, T. N., Tolo, A., Florez, T., & El Masri, Y. (2013). *Balancing trust and accountability? The Assessment for Learning programme in Norway.* Retrieved from: http://www.oecd.org/education/ceri/Norwegian%20GCES%20case%20study%20OECD.pdf. Paris: OECD.

Hopmann, S. T. (2003). On the evaluation of curriculum reforms. *Journal of Curriculum Studies*, *35*(4), 459–478. https://doi.org/10.1080/00220270305520

Isaacs, T. (2010). Educational assessment in England. *Assessment in Education: Principles, Policy & Practice*, *17*(3), 315–334.

Jarning, H., & Aas, G. H. (2008). Between common schooling and the academe: The International Examination Inquiry in Norway, 1935–1961. In M. Lawn (Ed.), *An Atlantic crossing? The work of the International Examination Inquiry, its researchers, methods and influence* (pp. 181–204). Oxford, UK: Symposium Books.

Jones, P. W. (2004). Taking the credit: Financing and policy linkages in the education portfolio of the World Bank. In G. Steiner-Khamsi (Ed.), *The global politics of educational borrowing and lending* (pp. 188–200). New York, NY: Teachers College Press.

Jonsson, A., Lundahl, C., & Holmgren, A. (2015). Evaluating a large-scale implementation of Assessment for Learning in Sweden. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 104–121. http://doi.org/10.1080/0969594X.2014.970612

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. http://doi.org/10.1016/j.edurev.2007.05.002

Kamens, D. H. (2015). Globalisation and the emergence of an audit culture: PISA and the search for "best practices" and magic bullets. In H. D. Meyer & A. Benavot (Eds.), *PISA, power, and policy: The emergence of global educational governance* (pp. 117–139). Oxford, UK: Symposium Books.

Kane, M. T. (1990). *An argument-based approach to validation.* ACT Research Report Series. Iowa City, IA: American College Testing Program. Retrieved from http://files.eric.ed.gov/fulltext/ED336428.pdf

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education / Praeger.

Kane, M. T. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182.

Kane, M. T. (2016). Explicating validity. Assessment in Education: Principles, Policy & Practice, 23(2), 198–211.

Kellaghan, T. (2001). The globalisation of assessment in the 20th century. *Assessment in Education: Principles, Policy & Practice*, *8*(1), 87–102.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37.

Kirton, A., Hallam, S., Peffers, J., Robertson, P., & Stobart, G. (2007). Revolution, evolution or a Trojan horse? Piloting assessment for learning in some Scottish primary schools. *British Educational Research Journal*, *33*(4), 605–627. http://doi.org/10.1080/01411920701434136

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284.

Kommunerevisjonen. (2009). *Avgangskarakterer i grunnskolen: Likebehandles elevene i grunnskolen?Report 9/2009* [Exit grades in compulsory education: Are the students treated equally? Report 9/2009]. Oslo, Norway: The City of Oslo:

Kommunerevisjonen. (2013). *Standpunktkarakterer i videregående skole: likebehandles elevene?* Rapport 17/2013 [Overall achievement grades in upper secondary education: Are the students treated equally? Report 17/2013]. Oslo, Norway: The City of Oslo.

Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative research interviewing*. Los Angeles, CA: Sage Publications.

Lau, A. M. S. (2016). "Formative good, summative bad?" A review of the dichotomy in assessment literature. *Journal of Further and Higher Education*, *40*(4), 509–525.

Lawn, M. (Ed.) (2008). The work of the International Examination Inquiry, its researchers, methods and influence. Oxford, UK: Symposium Books.

Lawn, M., & Grek, S. (2012). *Europeanizing education: Governing a new policy space.* Oxford, UK: Symposium Books.

Lindblad, S., Pettersson, D., & Popkewitz, T. S. (2015). *International comparisons of school results: A systematic review of research on large scale assessments in education.* Sub-project, part of the SKOLFORSK project. Stockholm, Sweden: The Swedish Research Council. Retrieved from https://publikationer.vr.se/produkt/international-comparisons-of-school-results-a-systematic-review-of-research-on-large-scale-assessments-in-education/

Littig, B. (2009). Interviewing the Elite – Interviewing Experts: Is There a Difference? In A. Bogner, B. Littig, & W. Menz (Eds.), *Interviewing experts* (pp. 98–116). London, UK: Palgrave Macmillan.

Luhmann, N. (1985). *A sociological theory of law*. London, UK: Routledge.

Luhmann, N., & Schorr, K. E. (1979/2000). *Problems of reflection in the system of education.* Münster, Germany: Waxmann.

Lundahl, C. (2006). *Viljan at veta vad andra vet* [The desire to know what other knows] (Doctoral dissertation). Uppsala University, Sweden.

Lundahl, C. (2008). Inter/national assessments as national Curriculum: The case of Sweden. In M. Lawn (ed.), *An Atlantic crossing? The work of the international examination inquiry, its researchers, methods and influence* (pp, 157–180). Oxford, UK: Symposium Books.

Lundahl, C. (2019). Making testers out of teachers: The work of a Swedish research institute. *History of Education*, doi: 10.1080/0046760X.2019.1565422

Lundahl C., Hultén M., Klapp, A., and Mickwitz, L. (2015) Betygens geografi - forskning om betyg och summa- tiva bedömningar i Sverige och internationellt [The Geography of grading - Research review on grading and summative assessments in Sweden and internationally.] Stockholm: Swedish Research Council.

Lundahl, C., & Tveit, S. (2014). Att legitimera nationella prov i Sverige och i Norge: en fråga om profession och tradition [To Legitimate national tests in Sweden and Norway: a question of profession and tradition]. *Pedagogisk forskning i Sverige, 19*(4–5), 297–323.

Lundahl, C., & Waldow, F. (2009). Standardisation and "quick languages": The shape-shifting of standardized measurement of pupil achievement in Sweden and Germany. *Comparative Education, 45*(3), 365–385.

Lundgren, U. P. (2011). PISA as a political instrument: One history behind the formulating of the PISA programme. In M.-A. Pereyra, H.-G. Kothoff, & R. Cowen (Eds.), *PISA under examination: Changing knowledge, changing tests, changing schools* (pp. 17–30). Rotterdam, the Netherlands: Sense Publisher.

Lysne, A. (2006). Assessment theory and practice of students' outcomes in the Nordic countries. Scandinavian Journal of Educational Research, 50, 327–359.

Marginson, S., & Mollis, M. (2001). "The door opens and the tiger leaps": Theories and Reflexivities of Comparative Education for a Global Millennium. *Comparative Education Review*, *45*(4), 581–615. https://doi.org/10.1086/447693

Marozzi, M. (2015). Measuring trust in European public institutions. *Social Indicators Research*, *123*(3), 879–895.

Mausethagen, S. (2013a). Reshaping teacher professionalism: An analysis of how teachers construct and negotiate professionalism under increasing accountability (Doctoral dissertation). Oslo and Akershus University College of Applied Sciences, Norway. Retrieved from https://oda-hioa.archive.knowledgearc.net/handle/10642/2922

Mausethagen, S. (2013b). Accountable for what and to whom? Changing representations and new legitimation discourses among teachers under increased external control. *Journal of Educational Change, 14*(4), 423-444. http://doi.org/10.1007/s10833-013-9212-y

Mausethagen, S. (2013c). Governance through concepts: The OECD and the construction of "competence" in Norwegian education policy. *Berkeley Review of Education*, *4*. doi:10.5070/B84110058

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). Washington, DC: American Council on Education.

Messick, S. (2005). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(5), 5–8.

Meyer, J. W., & Ramirez, F. O. (2003). The world institutionalization of education. In J. Schriewer (Ed.), *Discourse formation in comparative education* (pp. 111–132). Frankfurt am Main, Germany: Peter Lang.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, *83*(2), 340–363.

Mølstad, C. E. (2015a). State-based curriculum-making: Approaches to local curriculum work in Norway and Finland. *Journal of Curriculum Studies*, *47*, 441–461.

Mølstad, C. E. (2015b). State-based curriculum-making: A study of curriculum in Norway and Finland (Doctoral dissertation). University of Oslo, Norway.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. http://doi.org/10.1080/09695940701478321

Newton, P. E., & Shaw, S. (2014). Validity in educational and psychological assessment. London, UK: SAGE.

Nordin, A., & Sundberg, D. (2014). The making and governing of knowledge in the education policy field. In A. Nordin & D. Sundberg, *Transnational policy flows in European education* (pp. 9–20). Oxford, UK: Symposium Books.

Novak, J. (2018). *Juridification of Educational Spheres. The Case of Swedish School Inspection.* Uppsala: Acta Universitatis Uppsaliensis. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-335578

Novak, J., & Carlbaum, S. (2017). Juridification of examination systems: Extending state level authority over teacher assessments through regrading of national tests. *Journal of Educational Policy, 32*(5), 673–693. https://doi.org/10.1080/02680939.2017.1318454

OECD. (2005). Formative assessment: Improving learning in secondary classrooms. Paris, France: OECD.

OECD. (2013). *Synergies for better learning: An international perspective on evaluation and assessment.* OECD reviews of evaluation and assessment in education. Paris, France: OECD.

OECD. (2015). *Improving schools in Sweden: An OECD perspective.* Paris, France: OECD. Retrieved from http://www.oecd.org/edu/school/improving-schools-in-sweden-an-oecd-perspective.htm

Ozga, J., Dahler-Larsen, P., Segerholm, C., & Simola, H. (2011). *Fabricating quality in education: Data and governance in Europe*. London, UK: Routledge.

Parsons, T. (1960). A sociological approach to the theory of organization. In T. Parsons (Ed.), *Structure and Process in Modern Societies (pp. 16-58)*. Glencoe, IL: Free Press. (Originally published in 1956).

Pettersson, D. (2008). *Internationell kunskapsbedömning som inslag i nationell styrning av skolan* [International knowledge assessment as part of national governing of education] (Doctoral dissertation). Uppsala University, Sweden.

Pettersson, D., Prøitz, T. S., & Forsberg, E. (2017). From role models to nations in need of advice: Norway and Sweden under the OECD's magnifying glass. *Journal of Education Policy*, *32*(6), 721–744.

Phillips, D. (2004). Towards a theory of policy attraction in education. In G. Steiner-Khamsi (Ed.), *The global politics of educational borrowing and lending* (pp. 54–67). New York, NY: Teachers College Press.

Phillips, D., & Ochs, K. (2003). Processes of policy borrowing in education: Some explanatory and analytical devices. *Comparative Education Review*, *39*(4), 451–461.

Popham, W. J. (2011). *Classroom assessment: What teachers need to know* (6th ed.). Boston, MA: Pearson.

Prøitz, T. S. (2010). Learning outcomes: What are they? Who defines them? When and where are they defined? *Educational Assessment, Evaluation and Accountability, 22*(2), 119–137. http://doi.org/10.1007/s11092-010-9097-8

Prøitz, T. S. (2013). Variations in grading practice: Subjects matter. *Education Inquiry, 4*, 555-575. https://doi.org/10.3402/edui.v4i3.22629

Prøitz, T. S. (2015a). Uploading, downloading and uploading again: Concepts for policy integration in education research. *Nordic Journal of Studies in Educational Policy*. Retrieved from https://www.tandfonline.com/doi/abs/10.3402/nstep.v1.27015

Prøitz, T. S. (2015b). Learning outcomes as a key concept in policy documents throughout policy changes. *Scandinavian Journal of Educational Research, 59*(3), 275-296. https://doi.org/10.1080/00313831.2014.904418

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28,* 4-13.

Ramirez, F. O., Schofer, E., & Meyer, J. W. (2018). International Tests, National Assessments, and Educational Development (1970–2012). *Comparative Education Review*, *62*(3), 344–364. https://doi.org/10.1086/698326

Ringarp, J. (2011). Professionens problematik. Lärarkårens kommunalisering och valfärdsstatens förvandling [The professions problem. The teaching professions decentralisation and the changed welfare state]. Stockholm, Sweden: Makadam.

Ringarp, J., & Waldow, F. (2016). From "silent borrowing" to the international argument: Legitimating Swedish educational policy from 1945 to the present day. *Nordic Journal of Studies in Educational Policy,* 2016(1). https://doi.org/10.3402/nstep.v2.29583

Rhodes, R. (1997). Understanding governance: Policy networks, governance, reflexivity and accountability. Buckingham, United Kingdom: Open University Press

Rhodes, R. A. W. (2007). Understanding Governance: Ten Years On. *Organization Studies*, *28*(8), 1243–1264. https://doi.org/10.1177/0170840607076586

Robertson, S. L. (2005). Re-imagining and rescripting the future of education: Global knowledge economy discourses and the challenge to education systems. *Comparative Education Review*, *41*(2), 151–170. https://doi.org/10.1080/03050060500150922

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. https://doi.org/10.1007/BF00117714

Sahlberg, P. (2016). The global educational reform movement and its impact on schooling. In: K. Mundy, A. Green, B. Lingard, & A. Verger (Eds.), *Handbook of global education policy* (pp. 128–144). Hoboken, NJ: John Wiley & Sons.

Schreier, M. (2012). *Qualitative content analysis*. London, UK: Sage.

Schriewer, J. (1988). The method of comparison and the need for externalization: Methodological criteria and sociological concepts. In J. Schriewer & B. Holmes (Eds.), *Theories and methods in comparative education*. Frankfurt am Main, Germany: Peter Lang.

Schriewer, J. (1999). Coping with complexity in comparative methodology: Issues of social causation and processes of macro-historical globalization. In R. Alexander, P. Broadfoot, & D. Phillips (Eds.), *Learning from comparing. New directions in comparative educational research* (pp. 33–72). Oxford, UK: Symposium.

Schriewer, J. (2003). Comparative education methodology in transition: Towards the study of complexity? In J. Schriewer (Ed.), *Discourse formation in comparative education* (pp. 3–52). Frankfurt am Main, Germany: Peter Lang.

Schriewer, J. (2004). Multiple internationalities: The emergence of a world-level ideology and the persistence of idiosyncratic world-views. In C. Charle, J. Schriewer, & P. Wagner (Eds.), *Transnational intellectual networks* (pp. 473–533). Frankfurt am Main, Germany: Campus Verlag.

Schriewer, J. (2014). Neither orthodoxy nor randomness: Differing logics of conducting comparative and international studies in education. *Comparative Education*, *50*(1), 84–101. https://doi.org/10.1080/03050068.2014.883745

Scott, W. R. (2008). Institutions and organizations: Ideas and interests. Los Angeles, CA: Sage.

Scriven, M. (1967). The methodology of evaluation. In R. E. Stake (Ed.), *Curriculum evaluation* (pp. 39-83). American Educational Research Association (monograph series on evaluation, no. 1). Chicago, IL: Rand McNally.

Scriven, M. (1990). Beyond formative and summative evaluation. In K. Rehage, M. McLaughlin & D. Phillips (Eds.), *Evaluation and education: At quarter century. NSSE Yearbook.* Chicago: NSSE.

Silova, I. (2009). Varieties of educational transformation: The post-socialist states of Central/Southeastern Europe and the former Soviet Union. In R. Cowen & A. M. Kazamias (Eds.), *International handbook of comparative education* (pp. 295–420). London, UK: Springer.

Sjøberg, S. (2014). PISA-syndromet: Hvordan norsk skolepolitikk blir styrt av OECD [The PISA syndrome: How Norwegian school politics is governed by the OECD]. *Nytt Norsk Tidsskrift*, *31*(1), 30–43.

Skedsmo, G. (2011). Formulation and realisation of evaluation policy: Inconsistencies and problematic issues. *Educational Assessment, Evaluation and Accountability*, *23*(1), 5–20. https://doi.org/10.1007/s11092-010-9110-2

SOU (Official Government Report). (2016: 25). *Likvärdigt, rättssäkert och effektivt: ett nytt nationellt system för kunskapsbedömning* [Equal, legal, and effective – A new national knowledge assessment system]. Stockholm: Fritzes.

Spillane, J., & Burch, P. (2006). The institutional environment and instructional practice: Changing patterns of guidance and control in public education. In H.-D. Meyer & B. Rowan (Eds.), *New institutionalism in education* (pp. 87–102). Ithaca, NY: State University of New York Press.

Steiner-Khamsi, G. (2004). *The global politics of educational borrowing and lending*. New York, NY: Teachers College Press.

Steiner-Khamsi, G. (2010). The politics and economics of comparison. *Comparative Education Review,* 54(3), 323–342.

Steiner-Khamsi, G. (2014). Cross-national policy borrowing: Understanding reception and translation. *Asia Pacific Journal of Education, 34*(2), 153–167. doi:10.1080/02188791.2013.875649

Steiner-Khamsi, G., & Waldow, F. (Eds.) (2012). *Policy borrowing and lending. World Yearbook of Education 2012.* New York, NY: Routledge.

Stobart, G. (2008). Testing time. The uses and abuses of assessment. London, UK: Routledge.

Sundberg, D., & Wahlström, N. (2012). Standards-based curricula in a denationalised conception of education: The case of Sweden. *European Educational Research Journal*, *11*(3), 342–356.

Taras, M. (2005). Assessment – Summative and Formative – Some Theoretical Reflections. *British Journal of Educational Studies*, *53*(4), 466–478. https://doi.org/10.1111/j.1467-8527.2005.00307.x

Taras, M. (2007). Assessment for learning: understanding theory to improve practice. *Journal of Further and Higher Education*, *31*(4), 363–371. https://doi.org/10.1080/03098770701625746

Taras, M. (2009). Summative assessment: The missing link for formative assessment. *Journal of Further and Higher Education*, *33*(1), 57–69. http://doi.org/10.1080/03098770802638671

The National Commission on Excellence in Education. (1983). *A Nation at Risk. A Report to the Nation and the Secretary of Education United States Department of Education*. Retrieved from https://www.edreform.com/wp-content/uploads/2013/02/A_Nation_At_Risk_1983.pdf

Tønnessen, L. K. B., & Telhaug, A. O. (1996). *Elevvurderingen i norsk skole i etterkrigstida [Student assessment in Norwegian school post World War II].* In K. Skagen (Ed.), Karakterboka. Om karakterer og vurdering i ny skole [The marking book. About marks and assessment in new school] (pp. 15–34). Oslo: Universitetsforlaget.

Vlachos, J. (2018). *Trust-based evaluation in a market-oriented school system.* Retrieved from https://www.ifn.se/publikationer/working_papers/2018/1217

Vogt, B. (2017). Just assessment in school. A context-sensitive comparative study of pupils' conception in Sweden and Germany (Doctoral dissertation). Linnaeus University, Sweden

Wagemaker, H. (2013). International large-scale assessments: From research to policy. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 26–33). Boca Ration, FL: CRC Press.

Wahlström, N., & Sundberg, D. (2015). *En teoribaserad utvärdering av läroplanen Lgr 11* [A theory based evaluation of the curriculum Lgr 11], Uppsala: IFAU-Institutet för arbetsmarknads-och utbildningspolitisk utvärdering.

Waldow, F. (2009). Undeclared imports: Silent borrowing in educational policy-making and research in Sweden. *Comparative Education, 45*(4), 477–494.

Waldow, F. (2014). Conceptions of justice in the examination systems of England, Germany and Sweden: A look at safeguards of fair procedure and possibilities of appeal. *Comparative Education Review, 58*(2), 322–343.

Weber, M. (1922/1958). *Economy and society.* Selections reprinted in H. Gerth & C. W. Mills (Eds.), *From Max Weber*. New York, NY: Oxford University Press.

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, *21*(1), 1–19. http://doi.org/10.2307/2391875

Wermke, W. (2013). Development and autonomy: Conceptualising teachers' continuing professional development in different national contexts (Doctoral dissertation). Stockholm University, Sweden.

Wikström, C. (2005). Grade stability in a criterion-referenced grading system: The Swedish example. *Assessment in Education: Principles, Policy & Practice*, *12*(2), 125–144.

Wikström, C. (2006). Education and assessment in Sweden. *Assessment in Education: Principles, Policy & Practice*, *13*(1), 113–128. https://doi.org/10.1080/09695940600563470

Wikström, C., & Wikström, M. (2005). Grade inflation and school competition: An empirical analysis based on the Swedish upper secondary schools. *Economics of Education Review*, *24*(3), 309–322. http://doi.org/10.1016/j.econedurev.2004.04.010

Wiliam, D. (2008). Quality in Assessment. In S. Swaffield (Ed.), *Unlocking Assessment. Understanding for reflection and application* (pp. 123–137). New York, NY: Routledge.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14.

Ydesen, C. Ludvigsen, K., & Lundahl, C. (2013). Creating an educational testing profession in Norway, Sweden and Denmark, 1910–1960. *European Educational Research Journal, 12*(1), 120–138. https://doi.org/10.2304/eerj.2013.12.1.120

9 Appendices

Appendix 1: Policy documents analysed (sub-study III)

Appendix 2: Analyses of policy documents (sub study III)

Appendix 3: Interview guide example from the Norway case

Appendix 4: Interview guide example from the Sweden case

Appendix 5: Ethical approval documentation

Appendix 6: Five types of formative and summative assessment distinctions

9.1 Appendix 1: Policy documents analysed (sub-study III)

COUNTRY	NORWAY	SWEDEN
Ministry of Education and Research	NOU2002:10 (2002-06-14) NOU2003:16 (2003-06-05) Report to the Parliament 1 (2003–2004) Annex 3 (2004-04-02) Report to the Parliament 30 (2003–2004) (2004-04-02) Report to the Parliament 16 (2006–2007) (2006-12-15) Report to the Parliament 31 (2008–2009) (2008-06-13) Report to the Parliament 19 (2009–2010) (2010-06-11) Report to the Parliament 22 (2010–2011) (2011-04-29) Report to the Parliament 20 (2012–2013) (2013-03-14)	Ministry of Education (2004-12-22; 2005; 2008; 2008-09-25; 2009; 2011-11-24) Proposition 2008-09: 87 (2008-12-04) Proposition 2009-10:219 (2010-09-02) Proposition 2011-12:1 SOU2007:28 (2007-04) SOU2014:12 (2014-03) SOU2016:25 (2016-03) SOU2016:28 (2016-05) SOU2017:35 (2017-04) Swedish Parliament (2006)
Executive agency	DET (2010-12-22; 2016; 2017-02-21)	NAE (2000-08; 2001; 2002-09; 2004-02-23; 2007-06; 2008; 2009a; 2009b; 2011; 2012-10-19; 2013; 2014-06-12; 2015-04a; 2015-04b; 2015-11; 2017)

9.1.1 Norway – Ministry level policy documents

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.

Report to the Parliament 1 (2003–2004). *Stortingsmelding nr. 1 (2003-2004), tillegg 3.* [Report to the Parliament 1, Annex 3], pp. 1-6. Oslo: Ministry of Education and Research

Report to the Parliament 30 (2003–2004) (2004-04-02). *Kultur for læring* [Culture for learning]. Oslo: Ministry of Education and Research.

Report to the Parliament 16 (2006–2007). *... og ingen stod igjen. Tidlig innsats for livslang læring* [Early efforts for lifelong learning]. Oslo: Ministry of Education and Research

Report to the Parliament 31 (2007–2008). *Kvalitet i skolen* [Quality in schools]. Oslo: Ministry of Education and Research.

Report to the Parliament 19 (2009–2010). *Tid til læring* [Time for learning]. Oslo: Ministry of Education and Research.

Report to the Parliament 22 (2010–2011) (2011-04-29). *Motivasjon – mestring – muligheter* [Motivation, mastery – opportunities]. Oslo: Ministry of Education and Research.

Report to the Parliament 20 (2012–2013). *På rett vei* [On the right track]. Oslo: Norwegian Ministry of Education and Research.

Report to the Parliament 28 (2015–2016). *Fag – fordypning – forståelse – en fornyelse av kunnskapsløftet* [A renewal of the Knowledge Promotion]. Oslo: Norwegian Ministry of Education and Research

9.1.2 Norway: Executive Agency level policy documents

DET (2010). *Rammeverk for nasjonale prøver* [The National Testing Framework]. Oslo: Directorate for Education and Training

DET (2016): *Metodisk grunnlag for de nasjonale prøvene* [Methodological basis for the national tests]. Oslo: Directorate for Education and Training

DET (2017). *Rammeverk for nasjonale prøver* [The National Testing Framework]. Oslo: Directorate for Education and Training

9.1.3 Sweden: Ministry level policy documents

Bill 2008-09: 87. Tydligare mål och kunskapskrav – nya läroplaner för skolan [Clearer goals and knowledge requirements].

Bill 2009-10: 219. *Betyg från årskurs 6 i grundskolan* [Grades from Year 6].

Bill 2011/12:1. *Budgetpropositionen för 2012. Utbildning och universitetsforskning* [Budget bill for 2012. Education and university research].

Ministry of Education and Research (2004). *Uppdrag til Statens Skolverk avseende det nationella provsystemet* [Commission to the National Agency for Education regarding the national testing system]. Stockholm.

Ministry of Education and Research (2005). *Uppdrag til Statens Skolverk om nationella prov och diagnostisk stödmaterial [*Commission to the National Agency for Education regarding National Tests and Diagnostic Support Materials]. Stockholm.

Ministry of Education and Research (2008). *Regleringsbrev för budgetåret 2008 avseende Statens skolverk* [Regulation letter for the 2008 fiscal year regarding the National Agency for Education]. Stockholm.

Ministry of Education and Research (2008b). *Departementspromemoria. En individuell utvecklingsplan med skriftliga omdömen* [Ministry memorandum. More compulsory national tests in primary and lower secondary school]. Stockholm

Ministry of Education and Research (2009). *Uppdrag til Skolinspektionen om viss central rättning av nationella prov* [Commission to the Swedish School Inspectorate on certain re-marking of national tests]. Stockholm.

Ministry of Education (2011). *Uppdrag om nationella prov* [Commission regarding national tests]. Stockholm.

SOU (Official Government Report) 2007:28. *Tydliga mål och kunskapskrav i grundskolan.* Förslag till nytt mål- och uppföljningssystem [Clear goals and knowledge requirements in primary and lower secondary schools. Proposal for a new system for goal monitoring and follow-up. Stockholm: Fritzes

SOU (Government Official Report) 2014:12. *Utvärdera för utveckling – om utvärdering av skolpolitiska reformer* [Evaluate for development – on evaluation of school policy reforms]. Stockholm: Fritzes

SOU (Government Official Report) 2016:25. *Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning* [Equal, legal, and effective – a new national knowledge assessment system]. Stockholm: Fritzes

SOU (Government Official Report) 2016:38. *Samling för skolan. Nationella målsättningar och utvecklingsområden för kunskap och likvärdighet. Delbetänkande.* [Gathering for the school. National objectives and development areas for knowledge and equivalence. Interim report]. Stockholm: Fritzes

SOU (Government Official Report) 2017:35. *Samling för skolan. Nationella målsättningar och utvecklingsområden för kunskap och likvärdighet. Slutbetänkande* [Gathering for the school – National strategy for knowledge and equivalence]. Stockholm: Fritzes

Swedish Parliament 2006:19. *Översyn av grundskolans mål- och uppföljningssystem m.m. Kommittédirektiv* [Review of primary and secondary schools' goal and monitoring systems, etc. Committee Directive]. Stockholm.

9.1.4 Sweden: Executive Agency level policy documents

NAE (2000). Nationella kvalitetsgranskningar [National quality audits]. Stockholm: National Agency for Education.

NAE (2001). *Bedömning och betygssätnning. Kommentarar med frågor och svar.* [Assessment and grading. Comments with questions and answers]. Stockholm: National Agency for Education.

NAE (2002) *Betygssättning vid 18 fristående skolor* [Grading practices in 18 independent schools]. Stockholm: National Agency for Education.

NAE (2004). *Handlingsplan för en rättssäker och likvärdig betygssättning* [Action plan for fair and equal grading]. Stockholm: National Agency for Education.

NAE (2007). Provbetyg – Slutbetyg – Likvärdig bedömning? En statistisk analys av sambandet mellan nationella prov och slutbetyg i grundskolans årskurs 9, 1998-2006 [Test grades – Final grades – Equality. A statistical analysis of the relationship between national test grades and final grades in Year 9, 1998–2006]. Stockholm: National Agency for Education.

NAE (2008). *Mål och nationella prov i årskurs 3* [Goals and national tests in Year 3]. Stockholm: National Agency for Education.

NAE (2009). *Förslag på hur det nationella provsystemet bör utvecklas och utformas* [Proposal for how the national testing system should be developed and designed]. Stockholm: National Agency for Education.

NAE (2009). *Nationella prov årskurs 3* [Nationel tests for Year 3]. Stockholm: National Agency for Education.

NAE (2011). *Kunnskapsbedömning i skolan* [Knowledge Assessment in School]. Stockholm: National Agency for Education.

NAE (2012). *Uppdrag om kvalitetssäkring av nationella prov* [Commissoned quality asurance of the national tests]. Stockholm: National Agency for Education.

NAE (2013). *Resultat från ämnesproven i grundskolan våren 2012* [National tests in basic schools spring 2012]. Stockholm: National Agency for Education.

NAE (2014). *Uppdrag om kvalitetssäkring av nationella prov* [Commissoned quality asurance of the national tests]. Stockholm: National Agency for Education.

NAE (2015). *Provbetygens stabilitet – Om nationella prov åk 9 1998-2002* [Test grades' stability. On national tests – Year 9 1998-2012. Stockholm: National Agency for Education.

NAE (2015b). *Provpoängens tillförlitlighet. Om nationella prov. Skolverkets aktuella analyser 2015* [National test credit reliability. On national testing. NAE's current analyses 2015]. Stockholm: National Agency for Education.

NAE (2015c). *Skolreformer i praktiken. Hur reformerna landade i grundskolans vardag 2011–2014. School reforms in practice* [School reforms in practice. How the reforms were implemented in everyday school life]. Stockholm: National Agency for Education.

NAE (2017). *Nationella prov* [Nationella prov]. Retrieved from https://www.skolverket.se/bedomning/nationella-prov. Stockholm: National Agency for Education.

9.2 Appendix 2: Analyses of policy documents (sub-study III)

The emphasis on roles of educational assessment in national testing policy documents in Norway (2000–2017)

Policy document	Text in the policy documents addressing the purposes of the national tests	Classification
NOU2002:10 (2002-06-14) First Class from Class 1	A main intention of these tests will be that all users and participants get information about the development of central areas (p. 27). Disadvantages . . . : Schools with grades and examinations will also have to relate to two control systems for the students' and apprentices' learning outcomes. However, it is possible in the long run to think of a system where this type of test can replace all or part of today's systems with grades and examinations (p. 28). Advantages . . . : These types of tests can provide a more accurate picture of students' and apprentices' learning outcomes in general and the development over time for the entire country (p. 28).	GOVERN (CERTIFY)
Report to the Parliament 1, Annex 3 [2003–2004] (2002-11-02)	The ministry believes the tests should cover various purposes. They should both provide decision makers at various levels within the education sector a basis for implementing necessary actions in the sector and offer students and parents a better basis for requiring and/or participating in the improvement of the education provisioned (p. 4).	GOVERN (SUPPORT)
NOU 2003:16 (2003-06-05) First in line	The tests should cover two purposes: to serve as a basis for school evaluation and to give the students feedback on subject attainment and learning outcomes (p. 226).	GOVERN SUPPORT
Report to the Parliament 30, 2003–2004 (2004-04-02)	The ministry perceives it to be important that changes in (student) attainment can be traced from one year to another. To contribute to this, the ministry will make sure that necessary tools for making the final assessment standards are available. The ministry is planning that the national tests starting from 2004 will be important tools in this regard (p. 40). The ministry will consider whether the lower secondary exit examinations incrementally can be replaced with national tests (p. 41).	CERTIFY
Report to the Parliament 16 (2006–2007) (2006-12-15)	The tests should give information to students, teachers, school leaders, parents, municipalities, regional authorities, and the national level as bases for targeted development measures. The tests will be held early in the fall in Years 5 and 8 (p. 78). The ministry expects schools and municipalities to use the results of the national tests for follow-up work (p. 78).	GOVERN SUPPORT

Report to the Parliament 31 (2007–2008) (2008-06-13)	The purpose of the tests is first and foremost to give national and local authorities as well as school leaders good information about students' attainment. The results should form a basis for schools' and municipalities' development efforts (p. 80).	GOVERN SUPPORT
Report to the Parliament 19 (2009–2010) (2010-06-11)	The tests should provide information to students, teachers, school leaders, parents, school owners, regional authorities, and the national level as bases for improvement and development efforts (p. 21). The student and parents should get feedback from teachers about the tests, and how the tests will be followed up in the teaching. The test results should also be topics for the student conferences (p. 22).	GOVERN SUPPORT
DET (2010-12-22). The National Testing Framework	The national tests should assess the extent to which the students' skills are in accordance with the curriculum's aims for the basic skills of arithmetics and reading in Norwegian and English, as they are integrated into the competence aims for LK06 after the 4th and 7th year. The samples shall provide information to students, teachers, school leaders, guardians, municipalities, regional authorities and the national level as the basis for improvement and development (p. 5)	GOVERN (SUPPORT)
Report to the Parliament 22 (2010–2011) (2011-04-29)	The main purpose of national tests is to determine the extent to which a school has succeeded in developing students' basic skills. Information from the tests should be used as basis for quality development in schools, municipalities and on regional and national level. Additionally, the test results can contribute to strengthening schools work with adapted teaching (pp. 64–65).	GOVERN SUPPORT
Report to the Parliament 20 (2012–2013) (2013-03-14)	The purpose of tests is to determine students' skills and provide information to students, teachers, school leaders, parents, municipalities, and the regional and national levels, which can form a basis for improvement and development (p. 151).	GOVERN (SUPPORT)
DET (2016). Methodological Basis for the National Tests	As of 2014, results from all national tests are based on the use of IRT ("Item Response Theory") calibration and scaling methods where we use a 2-parameter IRT model. With the new model it is also possible to integrate an anchor test that ensures that the same number at all times describes the same skill. This gives us a measurement instrument that allows us to say something about changes from one year to the next (p. 3).	GOVERN
Report to the Parliament 28 (2015–2016) (2016-04-15)	(No discussion of national tests' role, despite a chapter on "Educational assessment in subjects".)	-

DET (2017-02-21). National Testing Framework	The purpose of national tests is to provide the school with knowledge about the students' skills in reading, arithmetic's and English. The information from the tests should form the basis for formative assessment (*underveisvurdering*) and quality development at all levels in the school system (p. 2). National tests provide information about individual students, groups, stages and schools that teachers and school leaders need to undertake quality development. For the student, the results of national tests, in accordance with the provisions in Chapter 3, should be a tool in the learning process as a basis for adapted education and help the student increase his / her competence in subjects (Regulations to the Education Act chapter 3).	SUPPORT GOVERN

The emphasis on roles of educational assessment in national testing policy documents in Sweden (2000–2017)

Policy document	Text in the policy documents addressing the purposes of the national tests	Classification
NAE (2000-08). National Quality Audits 2000. Report number 190.	Criteria for the evaluation of the grading practices...: The national tests are used and the results are discussed and used as support for the grading (p. 132). As stated above, we have found that there is a certain amount of monitoring of grade statistics and results on diagnostic and national tests. However, on the other hand, in-depth analyses of the results and measures based on these are lacking (p. 158). At the school level, the management should take more active responsibility regarding the monitoring of the different teachers' basis for grading, the results on national tests and final grades and analysis of underlying causes (p. 170). The results of the investigation show that there are significant shortcomings, in terms of fair and equal grading. Both the state and the municipalities seem to have underestimated the complexity of the goal and knowledge-referenced grading system and the time and power that were, and are, necessary to implement it (p. 175-176).	CERTIFY GOVERN
NAE (2001). Assessment and Grading. Comments with answers and questions.	One purpose of the tests is to contribute to as uniform a basis as possible for assessment across the country; another is to concretise the governing documents that form the basis for the planning and implementation of school activities (p. 28).	CERTIFY GOVERN
NAE (2002-09). The Grading at 18 free-schools.	Tests as support for the grading: (…) The [national tests] were considered by the teachers at both primary and secondary schools as an important basis for "calibrating grades on a national level" ("*rikslikare*"), but also as a valuable material in conversation with the students about what knowledge is sought. However, there were also objections to the national tests, such as that they tended to be	CERTIFY GOVERN

	too steering for the grading and that they caused stress for some students (p. 15). Criteria for evaluating the grading practices...: The national tests are used, and the results are discussed and used, as support for the grading (p. 37)	
NAE (2004-02-23). Action plan for a fair and equal grading.	The national test's role in supporting grading is improved to increase the possibilities for comparing test results between students and schools, and to ensure that test results can be more clearly discussed in relation to the final grades. To strengthen the students' situation, the NAE proposes that the students' perception of the extent to which they receive fair grades should always be reported in the quality reports (p. 2). A principal purpose of the national tests is to assist teachers in their grading of individual students (p. 3).	CERTIFY
Ministry of Education (2004-12-22). Commission to the National Agency for Education regarding the national testing system	The purpose should be to: Contribute to increased goal achievement for the students Clarify the goals and identify the strengths and weaknesses of students (diagnostic function) Specify course objectives and grading criteria Support an equal and fair assessment and grading Provide the basis for an analysis of the extent to which the knowledge objectives are reached at the school, municipality, and national levels (p. 1).	SUPPORT GOVERN CERTIFY
Ministry of Education (2005). Commission to the National Agency for Education regarding National Tests and Diagnostic Support Materials	The NAE shall, in its continued development: intensify the development of diagnostic support materials for early ages in primary school, especially in the field of reading development and mathematics, collaborate with the School Development Agency in order to increase the schools' knowledge and use of diagnostic support materials, review the national tests and the diagnostic materials for primary and lower secondary schools, as well as the national tests for upper secondary schools, and the tests' guidelines so that they do not disadvantage students based on aspects such as gender, ethnicity, and social background (p. 2).	SUPPORT
Swedish Parliament (2006). Committee Directive. Review of primary and secondary schools' goal and monitoring systems, etc.	The National Agency's system of national tests, diagnostic instruments, and commentary material is an important part of the monitoring system (p. 8). [The Time Table Delegation's] impression is that schools and principals do not make sufficient use of the results of the national tests to analyse the schools' goal attainment. The results are used primarily at the individual and national levels. This means that an important feature of the national tests is lost (p. 8). The national test system allows teachers to diagnose each student's knowledge development and supports equal assessment and grading (p. 9).	GOVERN (SUPPORT) (CERTIFY)
SOU2007:28. (2007-04). Clear goals and knowledge requirements in primary and lower	The municipalities' freedom to organise and pursue their school activities increased the demands for central follow-up and evaluation to determine that the school really provide all children with a common core of basic knowledge and, thus, reach the nationally established goals in this regard. At the local level, new	CERTIFY GOVERN

secondary school. Proposal for a new system for goal monitoring and follow-up.	timetables and curricula were envisaged to increase the needs of schools to verify that the education is on the right track and will meet national goals, which national tests can determine (p. 256). In the assignment letter to the Swedish NAE in 2004, two purposes were added: the test system should also contribute to the increased goal achievement for the students and to concretise course objectives and grade criteria. It can, thus, be noted that, over time, the government has increased the requirements for the test system to achieve and contribute (p. 261). That the tests shall also 'contribute to increased goal achievement for the students' was added as a new purpose, but it does not describe the way in which it was supposed to happen. An annex to the government decision shows that the national tests shall also continue to be "supportive of the grading, not steering as before" (p. 260). According to the NAE, the primary purpose of the Year 5 test is to assess the goal attainment. The tests are also said to have an important function by specifying the objectives of the curriculum and, to a limited extent, since they are given at the end of the current education period, can serve as diagnostic material (p. 266). It should be possible to use the national tests as support for grading in at least two functions: calibration and exemplification. The former function means that the tests should be an instrument that steers the grading. However, the extent to which the national tests are intended to determine the [final] grading must be considered. The provisions of the primary school regulation imply that the tests are to be used at the end of Year 9 and that they will be used to support the grading. The primary school regulation does not indicate the degree to which the tests shall affect the grading. This leads to high interpretation differences among teachers (p. 273). The second function of the test is that they can be exemplary models that show how national goals and knowledge requirements can be understood in relation to a subject content and student responses. This role has been highlighted by many teachers I have met during school visits. According to the teachers, the curriculum is perceived as being unclear and that it was not until test examples that they received support in their interpretation of the goals and the requirements for different grades (p. 273). The tests must be of such importance that they work as both guiding and steering the determination of final grades. This requires the curriculum to be designed to better steer both the teaching and the test construction. The national tests are assessed by using qualified methods and are, therefore, considered to be more accurate in testing students' knowledge in each subject than most of the teacher-made tests. It is not reasonable that a costly national testing system fails to steer the grading more than it does today (pp. 273-274). Since 2004, five purposes have been specified for the national testing system. The question is whether it can play all these roles simultaneously. My assessment is that the main task of the national tests is to support an equal assessment and a fair grading	

	process. The fact that the primary and lower secondary school regulation specifies how the tests shall be used in the grading supports my assessment (p. 280). The primary purpose of the test should not be to specify goals and knowledge requirements; they should support the teachers' assessment of the knowledge that the students demonstrate that they developed. It does not mean that the tests will not be able to function exemplarily in the future - assessment instructions with examples of how different knowledge qualities are assessed in relation to the knowledge requirements for different grade steps should continue to support teachers' interpretation of the target system – but it is not a primary purpose for the tests (p. 281). My assessment is, thus, that the testing system's current purposes should be limited and thereby clarified. A national test system should primarily: support an equal assessment of pupils' knowledge development and a fair grading process, as well as provide an analysis of the extent to which knowledge requirements are reached at the school, municipality, and national levels (p. 281).	
NAE (2007-06). Test grades – Final grades – Equality. A statistical analysis of the relationship between national tests grades and final grades in Year 9, 1998–2006.	The formulation in the primary school regulation (above) shows that what is strongest, when emphasised with regards to national tests, is the individual support for the teacher in the grading (p. 11). The national tests function and limitations. The collected test results are also intended to provide a basis for an analysis of the extent to which the knowledge goals are reached at the school, municipality, and national levels. Furthermore, they can provide a basis for analysing assessment and grading with an equivalence perspective, both at the local and national levels. However, this does not mean that the tests should have a steering effect in the sense that a teacher is expected to grade the students or class based on a certain relation to the national test results. The state has not expressed any perception of how close the final grades should be to the test grades – neither for individual students nor for classes, schools, or municipalities. There are, from the testing and assessment system's perspective, several good reasons for a teacher to assign a different final grade to an individual student than the result on the national test. A complete match between test grade and final grade for each individual student in a class or at a school would mean that the tests serve as examinations, which they are not supposed to do (p. 11). The NAE has, in several contexts, expressed strong concern about the tendency for national tests to be of such importance for the assessment of individual students that they have come to look like an examination test (…). National tests do not test all goals in the curriculum and they do not test all goals as much. It is simply too complicated to construct and execute a single test that would give a perfect picture of an individual's knowledge of a subject. The national tests are also not entirely reliable at an aggregated level as a measure of the knowledge of a subject (s. 11).	CERTIFY

Proposition 2008-09: 87. (2008-12-04). Clearer goals and knowledge requirements	However, the purposes of the test system have gradually been extended to include purposes such as specifying course objectives and grading criteria and contributing to increased goal attainment for students. In the government's opinion, the test system should primarily aim to support an equal assessment and fair grading, as well as provide the basis for analysing the extent to which knowledge requirements are reached at the school, municipality, and national levels (p. 19). The NAE studies show that in some schools there are systematic differences over time between national test results and final grades. The SSI also shows that assessment and grading is an area that has improved, even if major improvements are needed (p. 19). An analysis comparing national tests and final grades would probably be a valuable basis for the quality development of schools (p. 19). The government believes that the SSI, with tighter and sharper inspections, will contribute to the development and strengthening of the follow-up on both school and municipality levels (p. 19). According to the government, more and earlier mandatory national tests will strengthen the school's follow-up and evaluation of students' knowledge development. They will also strengthen the equivalence of teachers' assessment and grading. Ultimately, this is another step in the government's ambition to provide students with better prerequisites for increased goal attainment (p. 20).	CERTIFY GOVERN
Ministry of Education (2008-09-25). Regulation letter for the 2008 fiscal year regarding the National Agency for Education. Change decision 2008-09-25.	The NAE shall, in a readily accessible way, publish a statistical material where the difference between the final subject grade and the grades on national tests in Swedish, mathematics, and English in year 9 is shown at the school level.	CERTIFY GOVERN
Ministry of Education (2008). Ministry memorandum. More compulsory national tests in primary and lower secondary school.	There are several purposes for introducing more compulsory national tests in primary and lower secondary school and corresponding school forms. The tests shall provide support in the teacher's work with the students' learning. Teachers' assessment and grading of student knowledge must be more equal. The follow-up of students' knowledge needs to be strengthened to ensure that students have the right to an equal education and to increase their goals. The introduction of compulsory national tests in Year 3, 5, and 9 also enables the follow-up of students' knowledge throughout primary and lower secondary education. The information, thus, provides a better overview and a clearer picture of knowledge outcomes in primary and lower secondary school and corresponding school forms. National tests are also an important tool for teachers, schools, and school principals in assessing the need for support efforts when it comes to developing activities in the direction of national goals (p. 4-5).	SUPPORT CERTIFY GOVERN
NAE (2008). Goals and national tests in Year 3. Information	The purposes of the national tests are: To contribute to improved attainment To specify the goals and identify students' strong and weak sides	SUPPORT GOVERN

report about the implementation of national tests in Year 3.	To support an equal and fair (likvärdig och rättvis) assessment To provide a basis for an analysis of the extent to which the goals are achieved (p. 2)	
NAE (2009). Reporting on government commission to make suggestions for how the national testing system should be developed and designed.	The purpose of the national testing system is: To contribute to improved attainment To specify the goals and identify students strong and weak sides To specify course aims and grading criteria To support an equal and fair (likvärdig och rättvis) assessment To provide a basis for the analysis of the extent to which the goals are achieved on the local and national levels (p. 4) The purpose of national tests, in addition to the function of supporting grading, may be to obtain results compilations that can be used in comparison with other groups in the school system. Such comparisons can be made at the school, municipality, and national levels or between different groups of students based on gender, ethnicity, class, etc. (p. 7). It is also possible to include feedback to students which, beyond the summarised assessment, also provide forward-oriented information about how to proceed with the students' knowledge development or how the teaching should respond to the determined outcome. A good example for this is the national tests in Year 3 and 5, where one fills in knowledge profiles for each student. The knowledge profile weighs the result of the test, together with other assessment materials that the teachers have. Based on the knowledge profile, teachers and students can then plan the subsequent teaching (p. 7).	SUPPORT CERTIFY GOVERN
NAE (2009). Report on the commissioned assignment for national tests for Year 3	It was emphasised that the test should: support teachers' assessment of students' goal attainment and determine how well each student has attained the goals provide a basis for a knowledge profile to support knowledge development towards the goals, be used in follow-up and evaluation at different levels be included in a natural way in the teaching and include tasks that are familiar to the students, in terms of format, content, and time-taking, consider the age and varied development of the students (p. 2).	SUPPORT GOVERN
Ministry of Education (2009). Commission to the Swedish School Inspectorate on certain re-marking of national tests	Two important purposes of the national tests are that they will provide a basis for teachers' grades in a subject and that they contribute to an equal assessment throughout the country (p. 2). The government attaches great importance to the national tests in support of the follow-up of students' knowledge, but also in its function of contributing to an equal grading throughout the country. It is of the utmost importance to ensure that the correcting of the samples is done in an equal manner. The SSI should, therefore, be assigned the task of conducting a certain central correction of national tests. The purpose of a central correction of national tests is primarily to support an equal assessment and grading of the tests throughout the country. The core correction implies a quality assurance of the test system by creating an opportunity to detect error assessments that compromise system reliability. The activity will generate a	CERTIFY GOVERN

	statistical basis on how the national tests are assessed and graded in different subjects, years, and schools (p. 3).	
Proposition 2009-10:219 (2010-09-02). Grades from Year 6	Together with a developmental conference and the individual written development plan with reviews, action programs, and national tests, the grades are important tools for following up and evaluating the students' knowledge, thereby giving each student support in due time (p. 12).	SUPPORT
Swedish National Audit Office (2011-06-09). Equal grades, equal knowledge? A follow-up of government governing towards equivalent grading in primary and lower secondary school.	Since the academic year 2009/10, the NAE annually compares the deviations between the test grade and the final grade. According to the NAE, these summaries do not provide a comprehensive picture of the grade's equivalence, but they offer an indication of possible shortcomings. The summaries show that there are schools that show unreasonably large differences between test grades and final grades, while other schools show minor differences. This indicates, according to the NAE, that the grading fails to sustain equality (pp. 20-21). Through interviews with NAE representatives, it has been found that the agency, for many years, have pointed out to the government that the national tests have far too many goals and purposes; one explanation is that it is difficult to determine what the relationship and the co-variation, between national tests and final grades should be (p. 31).	CERTIFY GOVERN
NAE (2011). Knowledge Assessment in School	The purposes of the national tests and the assessment are primarily to: Support an equal and fair assessment Provide a basis for an analysis of the extent to which the goals are achieved The national tests can also contribute to: Specifying the curriculum An improved goal attainment for students (p. 54)	CERTIFY GOVERN (SUPPORT)
Proposition 2011-12:1, Expences Area 16	For national tests to contribute to an equal and fair grading, they are required to be conducted in a legal manner and that the assessment is objective. Therefore, the government intends to instruct the NAE to assure the national tests' quality. The government has estimated that SEK 15 million will be allocated per year from 2013 and onwards (p. 46).	CERTIFY
Ministry of Education (2011-11-24). Commission regarding national tests	The NAE shall: quality assure national tests so that they can be better used in the future to assess the development of knowledge over time by ensuring that the tests maintain an even degree of difficulty so that it is possible to follow and compare the results of the national tests by different schools and municipalities (…) review the compilation of a student's results on partial tests for a test grade in order to ensure that knowledge development can be traced over time, even after changing the grading scale develop the reports of results on national tests so that they become more analytical and, thus, provide better support for the national follow-up of the school's results (e.g., that it includes a comparison within and between schools and municipalities, a comparison of background factors, the development of outcomes over time, links to other national and international follow-ups and evaluations, and conclusions; p. 1-2).	GOVERN

NAE (2012-10-19). Reporting on government commission to quality assure national tests	In the *Report on the Commisioned Quality Assurance of National Tests* (NAE, 2012), the agency identified that the main purpose of the national testing system is, on the one hand, to support an equal and fair assessment, and to provide a basis for an analysis of the extent to which the goals are achieved on the school, municipality, and national levels (p. 1). Thus, an extension of the purposes of the national tests in mathematics is required to enable the use of the results for trend analysis of good validity. This threatens the other purposes of the test (e.g., the purpose of supporting an equal and fair grading process). Thus, the national tests may be less likely to support what is required under the provisions of the Education Act, the Schools Regulation, the Upper Secondary Education Regulation, and the Adult Education Regulation (p. 17).	CERTIFY GOVERN
NAE (2013). National tests in basic school spring 2012	The main purpose of the national tests is to support an equal and fair assessment and grading approach. The results provide a basis that schools and municipalities can use to analyse to which extent the goals are achieved. The national tests can also contribute to specify the curriculum (p. 34).	CERTIFY GOVERN (SUPPORT)
SOU2014:12 (2014-03). Evaluate for development - on evaluation of school policy reforms	One objection to the national tests is that they are unstable over time. Therefore, each new test must contain a number of data, such as anchor data, that was included in previous tests. In particular the national tests do not meet the requirement to contain a sufficient number of data each year to cover the content of the current subject. As an alternative, sampling tests provide a structure can to provide better trend measurement. By allowing a limited number of students (a selection) to respond to a limited number of partly different, non-public questions, the content of the topic in question can be covered (p. 18). Therefore, we propose an investigation to be commissioned to co-design the trend measurement system through random surveys and to investigate how the future system for the evaluation of knowledge outcomes should be designed in its entirety. This is especially important for the national tests that currently have many different purposes. The investigation should focus on clarifying the purposes that can and should be linked to the system, which instruments should be linked to the respective purposes, and to what actors' information needs the different instruments should respond (p. 19).	GOVERN
NAE (2014-06-12). Reporting on government commission to quality assure national tests	The conclusion is that if it is desirable for the national tests in mathematics to generate results that can be meaningfully used for annual trend analyses of different areas of knowledge; such trend analyses should be clearly identified as a purpose for the national tests. The NEA, however, recommends that the national tests' current purposes be retained and not extended. If knowledge development in various areas of knowledge is to be monitored with great certainty over time, it is, therefore, more reasonable to investigate the possibility of developing a specific measuring instrument for that purpose (p. 2). The efforts that the agency have undertaken to improve the stability of the tests was based on the tests' two purposes. Based on the first purpose—to support teachers' assessment and	CERTIFY GOVERN

	grading—it is important that the tests cover as much of the curriculum as possible and that the tasks vary so that the students have the chance to demonstrate what they can do. To provide a better basis for the second purpose—to monitor the school, municipality, and national levels—a range of overarching actions have been undertaken (p. 13). The national tests' steering function and grading supporting purpose stand in conflict with being a trend-measuring instrument of international standards. However, even if the national tests cannot be used to the fullest extent for trend measures, it is naturally of great importance for the national tests to still be developed so that they are as comparable over time as possible, even if the tests remain broad in form and content (p. 25).	
NAE (2015-04a). Tests grades' stability. On national tests – Year 9, 1998-2012	The purpose of the national tests has varied slightly during the current period, but the main intentions have been for the tests to support teachers' grading and provide a basis for assessing the attainment of goals. That is, the tests will serve the teachers in their task of providing fair and equal grades, as well as providing results that may be used to assess and evaluate the outcome of the school's work at the overall school, municipality, and national levels, thus serving as consumer information in an increasingly competitive school system (p. 8).	CERTIFY GOVERN
NAE (2015-04b). National tests credit reliability. On national testing. NAE's current analyses 2015.	The national tests currently have two stated purposes, namely to: support an equal and fair assessment and grading provide a basis for analysing the extent to which knowledge requirements are met at the school, municipality, and the national levels (p. 19).	CERTIFY GOVERN
NAE (2015-11). School reforms in practice. How the reforms landed in everyday school life	With earlier grading (and national tests), this information would become accessible at an earlier point, thereby producing good knowledge development for the student. In the objectives context, the informative function is especially highlighted when it comes to students with difficulties because the grades are thought to contribute to this factor being pointed out by the principal and teacher (p. 47).	SUPPORT
SOU2016:25 (2016-03). Equal, legal, and effective – a new national knowledge assessment system	As stated on the NEA's website, there are two main purposes for the national tests. The first purpose is to support an equal and fair assessment and grading process. The second is to provide a basis for analysing to what extent the knowledge requirements are achieved at the school, municipality, and national levels. In addition to the purposes, the website shows that the national tests can also help to crystallise the curricula and students' increased achievement (p. 102). A pure purpose for every part of the knowledge assessment system In this section, the investigation provides proposals for a pure purpose for each part of the new national knowledge assessment system (i.e., for national tests, national assessment support (*de nationella bedömningsstöden*), and the national knowledge evaluation). The proposals suggest that: The national tests should have only one purpose, which should be to support the grading process. Today's other purpose for the	CERTIFY GOVERN

	tests—to provide the basis for an analysis of the extent to which knowledge requirements are met at different levels—should instead be carried by the national knowledge assessment at the national level, as well as the grading and local evaluation tests at the school and municipality levels. The purpose of the 'national assessment support' (*de nationella bedömningsstöden*) is to provide support, diagnostic or formative. The NAE shall provide the purposes of, and information about, the national tests and the national assessment support that is available. In addition, the NAE shall inform on the purpose of the national knowledge evaluation (p. 229). To strengthen an equal grading process and the legal rights of students, we propose that the national tests' purposes shall be to support the grading. Equality and legal rights are important because grades are so important for the individual, especially in Year 9 and upper secondary school, where admission to different courses often takes place based on the grades (p. 233). Today, the NAE informs (e.g., on its website) about the purpose of the national tests and the national assessment support. However, there are teachers who feel that the information about the purposes is unclear. When the investigation now proposes a pure purpose for the national tests and for the different types of assessment support, extensive information efforts are needed to establish the new purposes. Therefore, we consider that the NAE should inform the public about the national tests and the national assessment support available, as well as their purpose (p. 236). The NAE's information should focus on the different groups concerned by the changed purpose of the tests and assessment support. It may concern teachers who are the main users of the actual tests and support, but also municipalities and principals who, for example, use the test results to see if the tests meet their grading support function. In addition to the information on the purpose of the tests and assessment support, the NAE should also inform that it is inappropriate to use the tests for purposes other than those for which they were intended. Through our proposals, every part of the knowledge assessment system is assigned a clarified purpose. We separate the purposes of different parts and purge the current purposes. This scrutiny creates a robust and clear system. However, it is not possible to fully streamline a purpose. As mentioned earlier, it is, for example, difficult to control how the test results are used and whether the use follows the intended purpose (p. 236-237).	
SOU 2016:28 (2016-05). Gathering for the school. National objectives and development areas for knowledge and equivalence. Interim report.	The purpose of the national tests should be streamlined to support a fair and equal grading process (p. 24). It may also be reasonable to specify improvement steps in relation to the national tests for Year 3 (…). Improvement steps will potentially ensure that more students will reach the required level of all sub-tests. However, the tests are not considered sufficiently stable to be used for follow-up over time. The National Testing Inquiry (SOU 2016: 25) further states that students who meet the requirements of the tests in Year 3 may still need support and lack the skills required in later years. A greater variation in content and	CERTIFY GOVERN

	difficulty level that increases accuracy and the prognostic value is perceived to be desirable (p. 81). The OECD (2015a) emphasises that the development of a coherent monitoring and evaluation system is an important improvement area for Sweden. Today's system contains many essential components, but it does not constitute a coherent and integrated system for the various levels of responsibility, nor does it contain enough essential data. Above all, reliable results that reflect students' knowledge development are not available. Results on national tests and grades are reported at the school, municipality, and national levels, but they are not appropriate instruments to monitor results development over time. The shortcomings are especially important for quality development at the local level. Teachers, school leaders, and municipalities lack instruments that allow students to follow the results over time and allow comparisons with, for example, national results. There is also, as the OECD points out, generally insufficient habits in many schools' work with quality development to systematically use data that reflects students' knowledge development (p. 129). For a long time, the NAE has proposed an improved continuous performance monitoring approach. The proposal has recently been repeated by the report on improved results in primary school (SOU 2014: 12) and in the report entitled *Equal Knowledge Assessment in and of Swedish Schools - Problems and Opportunities* (Gustafsson et al., 2014a). The issue has now been examined by the National Testing Inquiry (SOU 2016: 25). The investigation was commissioned to review current national testing programmes and to submit proposals for a system for ongoing national trend evaluation over time. The investigation's proposal in the latter area shows that experimental activities with national knowledge assessment should be conducted to provide information on knowledge development over time at the national level. The implementation shall be conducted in a randomised and digital manner. The extension will be gradual to incrementally include subjects such as English, mathematics, Swedish, and natural and social orientation, as well as multidisciplinary competencies. Of the proposal, it is also apparent that…the evaluation can be used for monitoring at school and municipality level. This makes it possible to [implement] evaluation tests where, for example, teaching and support efforts can be assessed. Another possibility that the proposed system can create, for example, is the development of value-added measures for both the school and national levels (p. 130). **Commission recommendation:** Continued development efforts for knowledge assessment competence are necessary. The purpose of the national tests should be streamlined to support fair and equal grading. Adjustments to the grading scale may be considered in the long term (p. 155). The grading scale with more grades that was introduced with the new curricula sought to increase the clarity of information to pupils and guardians and to give the teachers the opportunity to	

	increase precision in assessing students' knowledge and the degree of goal attainment. Reducing the distance between grades will also encourage students to exert extra effort as more students can reach the closest higher grade. The Commission, like the OECD (2015a), considers that assessment is a central area of competence that is of great importance to students' learning. A high level of competence in the teaching profession in the assessment of knowledge, both formative and summative, is crucial for teachers to lead students' learning towards knowledge goals (Håkansson & Sundberg, 2012). Since 2011, there has been a very high demand for support for teachers' assessment work. Questions about grading and assessment are among the most common for the School Information Service, which annually answers about 120,000 questions from teachers and others. According to the TALIS survey (NAE 2014a), the assessment of knowledge is the area that Swedish teachers consider to be most important, in terms of professional development. More than every fourth primary education teacher reported a strong need for further education in the area, which was significantly more than the OECD average. The School Inspectorate has emphasised to the OECD that Swedish teachers have undeveloped assessment skills, especially regarding the continuous formative assessment that is so important for pupils' learning, as well as for the development of teaching. The aim of early retrieval and providing adequate support to students at risk of falling behind, which the OECD recommends (2015a), requires that teachers for these years have sufficient assessment skills. The NAE has produced commentary material to the knowledge requirements of all subjects in elementary school, as well as a large amount of assessment support of various types for both primary and secondary schools. Competence development efforts have also been implemented. This is fully in line with the OECD recommendations and the Commission agrees that continuing skills development efforts are necessary. The Commission also wishes to emphasise the responsibility of teacher education to provide prospective teachers with a qualified basis in the field of knowledge assessment (p. 156). In its report (SOU 2016: 25), the National Testing Inquiry has proposed several changes in the national testing system. The purpose of the test shall be renamed to support the assessment of grading to enhance equivalence and the legal rights of students. Furthermore, additional suggestions include undertaking experiments that assess digital tests, external assessments of student responses, and the process of ensuring anonymity of student responses. The stability of the tests over time should be improved and the relationship between national tests and grades ought to be made clearer. A continued production of national assessment support, which also should be digitalised, is proposed. Without considering specific aspects, the Commission considers that the orientation for the development of the national tests proposed is appropriate, given the importance of the tests for fair and equal grading (p. 157).	

	In summary, the school commission considers that the school's wellbeing and mental health is an area that requires increased attention. The Commission considers it positive that the National Testing Inquiry (SOU 2016: 25) proposed a reduction in the number of compulsory national tests in Year 9 and upper secondary school and that schools should facilitate the students' work situation during the testing periods (p. 161). The OECD (2015a) has emphasised the importance of developing a school culture that increasingly engages and motivates all students to perform at a high level, as well as ensures that all students learn basic skills at an early age. A number of changes that have been implemented, as well as new proposals presented by the government, aim to strengthen primary and lower-level support and break the pattern that implied that support needs were discovered and efforts were made only at the end of primary school. Examples of this are...national tests in year 3….The Commission considers that these changes are important and that they need to be followed up carefully, both in terms of implementation and the effects they have caused. In Section 5 on governing and responsibilities, the Commission has emphasised the importance of national school development programs, such as the area of students with special needs. It is especially important to improve and develop early special educational efforts for students with reading and writing difficulties (p. 163).	
SOU2017:35 (2017-04). Gathering for the school – National strategy for knowledge and equivalence	The Commission believes that it is important for the government to give high priority to the work of developing and digitising national tests and developing assessment support and knowledge evaluations to enable teachers to work more efficiently and to increase the level of equality and students' legal rights (p. 20) A coherent monitoring and evaluation system More reliable outcomes information about students' knowledge development, at the school, municipality, and national levels, is required than what is provided by the grades and national tests. It is important for the local follow-up that there are instruments that allow, for example, for the assessment of the effects of teaching and support efforts. For an informed school choice, it is of the utmost importance that there is access to information about a school's contribution to students' knowledge development through so-called value-added measurements. Such data can provide more accurate quality information than grades and test results. Information about how many children and students in Swedish schools are in need and receive some kind of additional adjustments and special needs measures, as well as what actions are undertaken and how they work, are missing. Knowledge about which programs and actions is needed for additional adjustments and special needs support is required (p. 30). Measuring the equivalence of a school system is complicated. Different measurements show different aspects of equivalence (NAE, 2012). Some common aspects used to measure equality are…:	CERTIFY GOVERN

	The differences between the school's results on national tests, validity values, or in results on various international surveys (p. 283). A lack of all measures of equivalence makes it difficult to determine an acceptable or reasonable level. This is why, above all, it is the trend (i.e., the development over time), that is important to scan. One of the most widely used and accepted indicators of equivalence is the degree of correlation between students' aspects, such as the parents' income, education level, and migration. NAE (2012d) previously found that the differences in student results between schools and municipalities have gradually increased, which has led to concerns about an increased impact of family background in students' school achievements. However, summaries of studies conducted do not indicate any increase (e.g., Sundén & Werin, 2016). Current research indicates, however, that this relationship has increased significantly over the past two decades in Sweden (Gustafsson & Yang Hansen, 2017). This picture appears to be more finegrained than whether or not the parent is highly educated with respect to the parents' educational background. Thus, if a more nuanced measure that indicates the level of education in more levels is used, an increase in the relationship appears (p. 284). A coherent monitoring and evaluation system **Commission's assessment:** The National Monitoring and Evaluation System must be supplemented on important points. Reliable measurements of students' knowledge development need to be developed. Information about students in need of support, support measures, and the impact of the efforts is needed. The monitoring of the national objectives proposed by the Commission should be carried out in the coming years as a distinctly defined work. As for the regular national monitoring and evaluation system, the Commission, as in the initial report, wishes to emphasise the need to address certain shortcomings, as the OECD (2015a) also addressed. More reliable performance information about students' knowledge development, at the national level as well as at the school and municipality levels, is required than what is provided by grades and national tests. It is important for the local monitoring that there are instruments that allow, for example, for the assessment of the effects of teaching and support efforts. For an informed school choice, it is of the utmost importance that there is access to information about a school's contribution to students' knowledge development through so-called value-added measurements. They can provide more accurate quality information than grades and test results. As can be seen from Chapter 9, information is also missing regarding how many children and students in Swedish schools need and receive some additional adaptations and special support measures, what actions are undertaken, and how they work. Knowledge about which programs and actions is needed for additional adaption and special need support is necessary (p. 311). It may also be reasonable to specify improvement steps in relation to the national tests in Year 3....Improvement steps could mean	

	that more students will reach the level of requirements in all sub-tests. However, the samples are not considered sufficiently stable to be used for monitoring over time. The National Testing Inquiry (SOU 2016: 25) further addresses that students who meet the requirements of the tests in Year 3 may still need support and lack the skills required in later years. A greater variation in content and difficulty levels that increases the accuracy and the prognostic value is perceived to be desirable (p. 409-410).	
NAE (2017). Obtained on May 29th 2017 from: https://www.skolverket.se/bedomning/nationella-prov	The purpose of the national tests is mainly to support an equivalent and just assessment and marking and to provide a basis for analysing to what extent the "knowledge requirements are achieved" on the school, municipality, and national levels. The national tests can also help to crystallise curricula and subject plans, and students' increased achievement. The tests are mainly summative The national tests have, above all, a summative function. This means that they will serve as a reference point at the end of a year or a course and identify what qualities the student has in his/her knowledge of the subjects/courses where the tests are conducted. The tests can be used formatively The national tests can also be used as part of the assessment for learning that is part of the teaching. The test results provide good information about the skills that constitute the strengths and skills of students through teaching needs to develop more. In this way, the tests even fill a formative function. The tests also provide a picture of how teaching has worked that, in turn, can provide ideas for how teaching can be developed.	CERTIFY GOVERN (SUPPORT)

9.3 Appendix 3: Interview guide example from the Norway case (executive agency assessment department director)

Bakgrunnsdata:

1) Presentasjon av informanten
   a) Kan du først kort presentere deg selv, med fokus på din kompetanse og dine arbeidsoppgaver i Utdanningsdirektoratet?
2) Arbeidserfaring
   a) Hva gjorde du før du kom til Utdanningsdirektoratet? Hvilke arbeidsoppgaver har du hatt tidligere (både i Utdanningsdirektoratet og tidligere arbeidsplass)?

Generelt om arbeidet med elevvurdering i skolen

3) Arbeidsmåter i forvaltningen av elevvurdering i norsk skole
   a) Kan du opplyse om hvordan arbeidsdelingen og prosessene er mellom deg og dine kolleger i Utdanningsdirektoratet, og overfor Kunnskapsdepartementet og politisk ledelse?
4) Kort oversikt over arbeidsområdet
   a) Kan du gi en kort oversikt over hovedlinjene i Utdanningsdirektoratets arbeid med elevvurdering i Kunnskapsløftet? Hva var bakgrunnen for de grepene som ble gjort?
      i) Kvalitetsutvalgets innstilling og stortingsmelding 30 (2003-2004)
      ii) implementeringen av Kunnskapsløftet
      iii) tiden etter implementeringen av Kunnskapsløftet?
5) Internasjonal påvirkning
   a) Kan du opplyse om i hvilken grad Utdanningsdirektoratet har sett til andre land, i arbeidet med elevvurdering i Kunnskapsløftet?
      i) I hvilken grad, og på hvilke områder knyttet til elevvurdering, har OECD sine anbefalinger vært betydningsfulle for Utdanningsdirektoratets arbeid?
6) Lover og forskrifter
   a) Kan du fortelle litt om Utdanningsdirektoratets arbeidet med nye vurderingsforskrifter, i forbindelse med og etter innføringen av det nye læreplanverket og hvorfor dette ble gjort?
7) Klarere regelverk
   a) Er det noe du vil trekke fram som særlig utfordrende i utdanningsmyndighetenes arbeid med å skape et klarere regelverk?
8) Lærerutdanning og -etterutdanning
   a) Kan du opplyse om Utdanningsdirektoratets arbeid med elevvurdering i forbindelse med lærerutdanningen og etterutdanningen av lærere?

i) På hvilken måte har Utdanningsdirektoratet vært involvert i Kunnskapsdepartements arbeid med ny lærerutdanning?
ii) Satsingen Vurdering for Læring kommer vi tilbake til.

Ekstern vurdering

9) Formålet med ekstern vurdering:
   a) Kan du opplyse hvordan Utdanningsdirektoratet definerer formålet med eksamen?
10) Sammenhengen mellom ekstern vurdering og lærernes vurdering
   a) Hvilke sammenhenger ser Utdanningsdirektoratet mellom eksamenssystemet og lærernes standpunktvurderinger?
11) Arbeidet med sentralt gitt eksamen
   a) Kan du opplyse om Utdanningsdirektoratets arbeid med sentralt gitt eksamen?
      i) Prosessen med å utarbeide eksamensoppgaver
      ii) Skolere sensorer
      iii) Gjennomføre sensur
      iv) Rapportere resultater
      v) Faglig innhold
      vi) Regelverk
12) Pålitelighet ved sentralt gitt eksamen:
   a) Hva slags oversikt har Utdanningsdirektoratet over påliteligheten ved sensorvurderingene – altså i hvilken grad sensorene er enige om måloppnåelsen på eksamensbesvarelser og i hvilken grad flere sensorpar ville komme til samme resultat?
13) Lokalt gitt eksamen:
   a) Hvilket ansvar har Utdanningsdirektoratet knyttet til lokalt gitte eksemaner og prøver?
14) Høring om muntlig eksamen
   a) Kan du opplyse om arbeidet med høringen på nye regler for lokalt gitt (muntlig) eksamen?
15) Pålitelighet ved lokalt gitt eksamen
   a) Hva slags oversikt har Utdanningsdirektoratet over påliteligheten ved de lokalgitte eksamenene?

Standpunktvurdering

16) Pålitelighet i standpunktvurderingene
   a) Erfarer Utdanningsdirektoratet at standpunktkarakterene norske lærere (både på ungdomstrinnet og i videregående skole) setter er sammenlignbare?
      i) Mellom klasserom, mellom skoler, mellom kommuner og fylker
17) Tiltak:
   a) Hvilke verktøy og virkemidler bruker Utdanningsdirektoratet for å ivareta dette hensynet?
      i) (Retningslinjer, veiledninger e.l.)

Nasjonale prøver

Nasjonale prøver har per i dag ikke en særlig funksjon som del av sluttvurdering og karaktersetting, men jeg vil likevel stille deg noen spørsmål om erfaringene med utviklingen av dette prøvesystemet.

18) Formål nasjonale prøver – og endringer av prøvene
    a) Kan du opplyse hvordan Utdanningsdirektoratet definerer nasjonale prøvers formål?
    b) Kan du opplyse om forarbeidet og implementeringen av de første nasjonale prøvene i 2004 og 2005, og de forandringene som siden har blitt gjort i rammeverket?

Kvalitetsutvalget foreslo i sin tid at nasjonale prøver på sikt kunne erstatte eksamen, og dette ble fulgt opp i stortingsmelding 30 og av stortingsflertallet.

19) Formålet om å erstatte eksamen
    a) Kan du opplyse hvordan Utdanningsdirektoratet arbeidet med for å ivareta dette formålet for prøvene, i den tidlige fasen av utviklingen av nasjonale prøver?
        i) Når og hvorfor forlot man ideen om at nasjonale prøver på sikt skulle erstatte eksamen?
        ii) Hvor langt kom man med å følge opp stortingsmeldingen og stortingets vedtak om dette?
20) Andre prøver
    a) Kan du opplyse om andre prøver Utdanningsdirektoratet har ansvar for, som har en funksjon i forhold til sluttvurdering?

Prosjekt Bedre vurderingspraksis og satsingen Vurdering for Læring

21) Vurdering FOR Læring
    a) Kan du opplyse om prosjektet Bedre Vurderingspraksis og den pågående satsingen Vurdering for Læring, og hvilke erfaringer Utdanningsdirektoratet har gjort seg med dette?
        i) Overordnet – rapportene forteller detaljer om gjennomføringen.
22) «Kjennetegn på måloppnåelse»
    a) Hva er bakgrunnen for at man utviklet begrepet «kjennetegn på måloppnåelse»?
        i) I internasjonal terminologi brukes begrepene kriterier og standarder.
        ii) Hvilke erfaringer gjorde man seg med veiledning for elevvurdering og eksempler på kjennetegn på måloppnåelse?
23) Nasjonale eller lokale kjennetegn på måloppnåelse
    a) I evalueringen av prosjektet anbefalte forskerne at det ble utviklet nasjonale veiledende vurderingskriterier. Utdanningsdirektoratet fulgte (med en del presiseringer) denne anbefaling i rapporteringen til Kunnskapsdepartementet på oppdraget, men Kunnskapsdepartementet har ikke gått inn for nasjonale kjennetegn på måloppnåelse. Kan du opplyse om de begrunnelsene Utdanningsdirektoratet har fått for ikke å innføre nasjonale bestemmelser om dette?
24) Kompetansemålene og læreplanene
    a) Kan du opplyse om prosessen med utarbeiding av læreplanene for Kunnskapsløftet, før implanteringen i 2006, og vurderingene som ble gjort med henhold til hvordan rammer for elevvurdering skulle tas hånd om i læreplanene?
        i) Hvilke dilemmaer står man overfor

ii) Hvilke konsekvenser har læreplanenes utforming fått for arbeidet med kjennetegn på måloppnåelse/vurderingskriterier siden?

25) Karakterskalaen
    a) Kan du opplyse om vurderingene som lå bak endringen av karakterskalaen i 2007?
26) Kan du opplyse de mest vesentlige endringene som ble gjort i vurderingsforskriftene i 2009?
    a) Presisering av grunnlaget for vurderingen
    b) Krav til dokumentasjon av underveisvurdering
    c) Forholdet mellom underveis- og sluttvurdering. Sluttvurdering underveis. Ikke gjennomsnitt.
27) Andre tiltak
    a) Er det andre tiltak Utdanningsdirektoratet planlegger å gjennomføre knyttet til regelverket, kjennetegn på måloppnåelse, sluttvurdering og karaktersetting?

Tre generelle temaer til slutt:

28) Om reformen:
    a) Erfarer Utdanningsdirektoratet at læreplanreformen Kunnskapsløftet gir et godt grunnlag for arbeid med elevvurdering?
29) Utdanningsprofesjonens kompetanse:
    a) Kan du opplyse om Utdanningsdirektoratets erfaringer med utdanningsprofesjonens (lærere, kommuner, fylkeskommuner, og nasjonale myndigheters) kompetanse knyttet til elevvurdering, fra forarbeidet til Kunnskapsløftet og fram til i dag?
30) Særlige utfordringer ved reformen:
    a) Er det noen sider ved reformen som utfordrer arbeidet med elevvurdering spesielt?
        i) Er det noen hensyn som har vært særlig vanskelige?
31) Sluttkommentar?
    a) Er det noe vi har snakket om som du gjerne ville sagt mer om, eller har du noen avsluttende kommentarer til Utdanningsdirektoratets arbeid med sluttvurdering i grunnopplæringen?

9.4 Appendix 4: Interview guide example from the Sweden case (executive agency testing and assessment unit director)

Bakgrunnsdata:

1. Kan du først kort presentere deg selv, med fokus på din kompetanse og dine arbeidsoppgaver i Skolverket?
2. Hvor lenge har du arbeidet i Skolverket, og hvilke arbeidsoppgaver har du hatt tidligere?

Generelt

3. Kan du gi en kort oversikt over Skolverkets arbeid med nationella prov, bedömning og betygsätting i skolen?
   a) Med fokus blant annet på hvordan arbeidet er organisert i prov- og bedömningsenheten
4. Hvordan arbeider Skolverket overfor Utbildningsdepartementet og andre offentlige instanser, i forbindelse med nationella prov og betygsätting?
   a) Eksempelvis når det kommer kritiske oppslag i pressen.

Nationella prov

5. Kan du opplyse om Skolverkets arbeid med de nationella provene i Sverige de seneste årene? Med fokus på:
   a. Syftet (formålet) med provene
   b. Omfang av provene (de enkelte prov og delprov)
   c. Emner/kurs med prov
   d. Provenes kvalitet
   e. Sensur av provene
   f. Pålitelighet i provbedömningen
   g. Kvalitetssikring (Skolinspektionen)
   h. Kostnader
6. Hvilket ansvar har Skolverket knyttet til andre prov enn de nationella?
7. Hva slags oversikt har Skoleverket over påliteligheten ved lærernes bedömninger utover de nationella provene?

Betygsätting

8. Kan du opplyse litt overordnet om arbeidet med ny betygskala i Sverige
9. Kan du opplyse om bakgrunnen for at man innførte flere betygnivåer (A-B-C-D-E)?
   a) Hva var bakgrunnen for kun å spesifisere A, C og E?
   b) I hvilken grad har det vært en diskusjon om å gjøre det såkalt «standards based», og dilemmaer knyttet til instrumentell tilnærming?
10. Kan du opplyse om bakgrunnen for at man har innført betygsätting fra årskurs 6?
    a) Hvilke hensyn har blitt veid opp mot hverandre da man besluttet tidligere betygsätting?
    b) Hvilke begrunnelser har det blitt lagt mest vekt på?

11. I hvilken grad erfarer Skolverket at betygene svenske lærere (både i grundskolan og i gymnaset) setter er sammenlignbare (på tvers av klasserom, skoler, kommuner)?
12. Kan du opplyse litt mer om Skolverkets studier av hvordan lærere setter betyg, i forhold till elevenes resultat på de nationella proven?
    a) Det har kommet en del påstander om at friskolene setter bedre betyg enn offentlige skoler, og det er forsket en del på dette. Hvordan vil du beskrive situasjonen omkring dette mulige problemet?

Gjennom årenes løp av inflasjon i lærernes betygsätting vært et tilbakevendende tema. Dette har særlig blitt dokumentert gjennom å sammenligne lærernes betygsätting med nationella provresultater.

13. Hva tror du denne inflasjonen skyldes?
    a) Stödinnsatser efter provet
    b) Betygsättning på andre grunder
    c) Lærere har svårt att underkänna elever
    d) Olika tolkningar av mål og kriterier

Slik jeg har oppfattet det er det i dag ingen retningslinjer på i hvilken grad lærerne skal vektlegge de resultatene på nationella prov i betygsättingen.

14. Har det vært slike retningslinjer tidligere?
    a) Har dette vært diskutert? (Fordeler/ulemper med en slik ordning)

Slik jeg har oppfattet det er det i dag ingen krav om «sambedömning» verken ved nationella prov eller annen betygsätting.

15. Er det riktig oppfattet?
    e) Har dette vært diskutert?

Norge har vi et eksamenssystem, der det er uavhengige lærere som sensurerer eksamenene, mens det i Sverige er elevenes egne lærere som gjør dette.

16. Har problemer knyttet til subjektiv bedömning av de nationella provene vært vurdert?

Noen generelle temaer til slutt:

17. Erfarer Skolverket at læreplanreformene (LGR 11 og GY 11) gir et godt grunnlag for arbeid med bedömning og betygsetting?
18. Er det noen sider ved reformene (LGR 11 og GY 11) som utfordrer arbeidet med, nationella prov, bedömning og betygsetting spesielt?
19. Kan du fortelle litt om i hvilken grad Skolverket har sett til andre land, i arbeidet med nationella prov og betygsätting i skolen?
    a) OECD? Norge? Danmark?
    b) Kjennskap til eksamenssystemene i Norge og Danmark? Vurdert i Sverige?
20. Er det noe vi har snakket om som du gjerne ville sagt mer om, eller har du noen avsluttende kommentarer til Skolverkets arbeid med bedömning og betygsätting i grunnopplæringen?

9.5 Appendix 5: Ethical approval documentation

Norsk samfunnsvitenskapelig datatjeneste AS
NORWEGIAN SOCIAL SCIENCE DATA SERVICES

MELDESKJEMA

Meldeskjema (versjon 1.4) for forsknings- og studentprosjekt som medfører meldeplikt eller konsesjonsplikt (jf. personopplysningsloven og helseregisterloven med forskrifter).

1. Prosjekttittel		
Tittel	Assessment for Selection in the Scandinavian Education Systems	
2. Behandlingsansvarlig institusjon		
Institusjon	Universitetet i Oslo	Velg den institusjonen du er tilknyttet. Alle nivå må oppgis. Ved studentprosjekt er det studentens tilknytning som er avgjørende. Dersom institusjonen ikke finnes på listen, vennligst ta kontakt med personvernombudet.
Avdeling/Fakultet	Det utdanningsvitenskapelige fakultet	
Institutt	Pedagogisk forskningsinstitutt	
3. Daglig ansvarlig (forsker, veileder, stipendiat)		
Fornavn	Sverre	Før opp navnet på den som har det daglige ansvaret for prosjektet.Veileder er vanligvis daglig ansvarlig ved studentprosjekt. Veileder og student må være tilknyttet samme institusjon. Dersom studenten har ekstern veileder, kan biveileder eller fagansvarlig ved studiestedet stå som daglig ansvarlig.Arbeidssted må være tilknyttet behandlingsansvarlig institusjon, f.eks. underavdeling, institutt etc. NB! Det er viktig at du oppgir en e-postadresse som brukes aktivt. Vennligst gi oss beskjed dersom den endres.
Etternavn	Tveit	
Akademisk grad	Høyere grad	
Stilling	Stipendiat	
Arbeidssted	Pedagogisk forskningsinstitutt	
Adresse (arb.sted)	Forskningsparken II plan 3, Gaustadalléen 21	
Postnr/sted (arb.sted)	0349 Oslo	
Telefon/mobil (arb.sted)	22850452 / 41545503	
E-post	sverre.tveit@ped.uio.no	
4. Student (master, bachelor)		
Studentprosjekt	Ja ○ Nei ●	
5. Formålet med prosjektet		
Formål	The ASSESS study investigates the rationales and policies for accreditation and selection of secondary students in the Scandinavian education systems; and the national authorities' strategies for ensuring validity and comparability of the assessments that underpin these accreditation and selection procedures.	Redegjør kort for prosjektets formål, problemstilling, forskningsspørsmål e.l. Maks 750 tegn.
6. Prosjektomfang		
Velg omfang	● Enkel institusjon ○ Nasjonalt samarbeidsprosjekt ○ Internasjonalt samarbeidsprosjekt	Med samarbeidsprosjekt menes prosjekt som gjennomføres av flere institusjoner samtidig, som har samme formål og hvor personopplysninger utveksles.
Oppgi øvrige institusjoner		
Oppgi hvordan samarbeidet foregår		
7. Utvalgsbeskrivelse		
Utvalget	Informanter i norsk, svensk og dansk utdanningsforvaltning	Med utvalg menes dem som deltar i undersøkelsen eller dem det innhentes opplysninger om. F.eks. et representativt utvalg av befolkningen, skoleelever med lese- og skrivevansker, pasienter, innsatte.
Rekruttering og trekking	Henvendelse til de aktuelle departementer, direktorater ol.	Beskriv hvordan utvalget trekkes eller rekrutteres og oppgi hvem som foretar den. Et utvalg kan trekkes fra registre som f.eks. Folkeregisteret, SSB-registre, pasientregistre, eller det kan rekrutteres gjennom f.eks. en bedrift, skole, idrettsmiljø, eget nettverk.

Førstegangskontakt	Førstegangskontakten opprtettes ved henvendelse på e-post enten til den aktuelle informanten selv, eller til deres overordnede eller saksbehandler.	Beskriv hvordan førstegangskontakten opprettes og oppgi hvem som foretar den. Les mer om dette på våre temasider.
Alder på utvalget	□ Barn (0-15 år) □ Ungdom (16-17 år) ■ Voksne (over 18 år)	
Antall personer som inngår i utvalget	15-20 personer (5-7 i hvert av landene Norge, Sverige og Danmark)	
Inkluderes det myndige personer med redusert eller manglende samtykkekompetanse?	Ja ○ Nei ●	Begrunn hvorfor det er nødvendig å inkludere myndige personer med redusert eller manglende samtykkekompetanse. Les mer om Pasienter, brukere og personer med redusert eller manglende samtykkekompetanse
Hvis ja, begrunn		
8. Metode for innsamling av personopplysninger		
Kryss av for hvilke datainnsamlingsmetoder og datakilder som vil benyttes	□ Spørreskjema ■ Personlig intervju □ Gruppeintervju □ Observasjon □ Psykologiske/pedagogiske tester □ Medisinske undersøkelser/tester □ Journaldata □ Registerdata □ Annen innsamlingsmetode	Personopplysninger kan innhentes direkte fra den registrerte f.eks. gjennom spørreskjema, intervju, tester, og/eller ulike journaler (f.eks. elevmapper, NAV, PPT, sykehus) og/eller registre (f.eks. Statistisk sentralbyrå, sentrale helseregistre).
Annen innsamlingsmetode, oppgi hvilken		
Kommentar		
9. Datamaterialets innhold		
Redegjør for hvilke opplysninger som samles inn	Opplysninger om arbeidet med elevvurdering siden forberedelsen av utdanningsreformen Kunnskapsløftet, og om hvilke utfordringer nasjonale utdanningsmyndigheter er opptatt av i det videre arbeidet med kvalitetsutvikling i grunnopplæringen knyttet til elevvurdering.	Spørreskjema, intervju-/temaguide, observasjonsbeskrivelse m.m. sendes inn sammen med meldeskjemaet. NB! Vedleggene lastes opp til sist i meldeskjema, se punkt 16 Vedlegg.
Samles det inn direkte personidentifiserende opplysninger?	Ja ○ Nei ●	Dersom det krysses av for ja her, se nærmere under punkt 11 Informasjonssikkerhet. Les mer om hva personopplysninger er NB! Selv om opplysningene er anonymiserte i oppgave/rapport, må det krysses av dersom direkte og/eller indirekte personidentifiserende opplysninger innhentes/registreres i forbindelse med prosjektet.
Hvis ja, hvilke?	□ 11-sifret fødselsnummer □ Navn, fødselsdato, adresse, e-postadresse og/eller telefonnummer	
Spesifiser hvilke		
Samles det inn indirekte personidentifiserende opplysninger?	Ja ● Nei ○	En person vil være indirekte identifiserbar dersom det er mulig å identifisere vedkommende gjennom bakgrunnsopplysninger som for eksempel bostedskommune eller arbeidsplass/skole kombinert med opplysninger som alder, kjønn, yrke, diagnose, etc. Kryss også av dersom ip-adresse registreres.
Hvis ja, hvilke?	Stilling på arbeidsplassen. Informantene blir informert om at de som kjenner arbeidsplassen deres vil kunne identifisere dem (dette er uunngåelig i denne typen undersøkelser der man benytter ekspertintervjuer av nøkkelinformanter).	
Samles det inn sensitive personopplysninger?	Ja ○ Nei ●	

Norwegian Centre for Research Data (NSD) approval the project and extended the permission to storage data Final report confirming that the data is anonymised was submitted to NSD on May 2nd 2018.

9.6 Appendix 6: Five types of formative and summative assessment distinctions

A: Definitions distinguishing between, timing, use and purposes		
Scholar	**Formative assessment**	**Summative assessment**
Scriven, 1967	It may have a role in the on-going improvement of the curriculum (…).	In another role, the evaluation process may serve to enable administrators to decide [the quality of] the entire finished curriculum (…)
Bloom et al (1971, p. 61)	to help both the learner and the teacher focus upon the particular learning necessary for movement towards mastery. (Interim test)	directed towards a much more general assessment of the degree to which the larger outcomes have been attained over the entire course or some substantial part of it (Final test)
Sadler (1989, p. 120)	Concerned with how judgments about the quality of student responses (performances, pieces, or works) can be used to shape and improve the student's competence by short-circuiting the randomness and inefficiency of trial-and-error learning.	Concerned with summing up or summarizing the achievement status of a student, and is geared towards reporting at the end of a course of study especially for purposes of certification.
B: Definitions without (explicit) summative assessment definitions		
Scholar	**Formative assessment**	**Summative assessment**
Black and Wiliam (1998a, p. 8)	All those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities'	[-]
Black & Wiliam (2009, p. 9)	Practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited.	[-]
C: Definitions which explicitly distinguishes between two types of summative assessment		
Scholar	**Formative assessment**	**Summative assessment**
Black (1998, p. 35)	1) Formative, to aid learning	2) Summative, for review, transfer and certification, and 3) Summative for accountability to the public

OECD (2005, p. 21)	Formative assessment refers to frequent, interactive assessments of students' progress and understanding to identify learning needs and adjust teaching appropriately.	Summative assessments are used to measure what students have learnt at the end of a unit, to promote students, to ensure they have met required standards on the way to earning certification for school completion or to enter certain occupations, or as a method for selecting students for entry into further education. Ministries or departments may use summative assessments and evaluations as a way to hold publicly funded schools accountable for providing quality education.
Stobart (2008, p. 24)	1) Selection and certification	2) Determining and raising standards 3) Formative assessment – assessment *for* learning
D: Definitions which consider summative assessment as foundational to formative assessment		
Scholar	**Formative assessment**	**Summative assessment**
Taras (2007, p. 367)	2) Subsequent to this judgement, it is possible to judge how improvements can be made (formative assessment).	1) Assessment is a judgement within agreed parameters: this judgement is a summation at any given point in time (summative assessment).
E: Summative assessment defined as a *judgment* and formative assessment as a way of using it		
Scholar	**Formative assessment**	**Summative assessment**
Newton (2007)	*FA is about a particular way of using information from summative judgments*	*A judgment*

Part II: The Articles

Article I

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

Submitted November, 2012; Accepted July 24, 2013; Published September, 2013.

Article II

Tveit, S., & Lundahl, C. (2018). New modes of policy legitimation in education: (Mis)using comparative data to effectuate assessment reform. *European Educational Research Journal, 17*(5), 631–655. https://doi.org/10.1177%2F1474904117728846

Submitted July, 2016; Accepted August, 2017; Published September, 2017.

Article III

Tveit, S. (2018a). Ambitious and ambiguous: Shifting purposes of national testing in the legitimation of assessment policies in Norway and Sweden (2000–2017). *Assessment in Education: Principles, Policy & Practice*, 25(3), 327–350. https://doi.org/10.1080/0969594X.2017.1421522

Submitted April, 2017; Accepted December, 2017; Published February, 2018.

Article IV

Tveit, S. (2018b). Transnational trends and cultures of educational assessment: Reception and resistance of national testing in Norway and Sweden During the Twentieth Century. In C. Alarcon & M. Lawn (Eds.), Assessment cultures. Studia Educationis Historica. Berlin, Germany: Peter Lang. https://doi.org/10.3726/978-3-653-06867-2

Submitted May 2016; Accepted August 2016; Published February 2018.

II

General submission

EERJ European Educational Research Journal

New modes of policy legitimation in education: (Mis)using comparative data to effectuate assessment reform

European Educational Research Journal
2018, Vol. 17(5) 631–655

DOI: 10.1177/1474904117728846
journals.sagepub.com/home/eer
SAGE

Sverre Tveit
Department of Education, University of Agder, Norway

Christian Lundahl
School of Humanities, Education and Social Sciences, Örebro University, Sweden

Abstract

Identifying three *modes of policy legitimation* in education, illustrated by shifts in Swedish educational assessment and grading policies over the past decades, the paper demonstrates significant trends with regard to national governments' policymaking and borrowing. We observe a shift away from *collaboracy* – defined as policy legitimation located in partnerships and networks of stakeholders, researchers and other experts – towards more use of supranational agencies (called *agency*), such as the Organisation for Economic Co-operation and Development, the European Union and associated networks, as well as the use of individual consultants and private enterprises (called *consultancy*) to legitimate policy change. Given their political and high-stakes character for stakeholders, assessment and grading policies are suitable areas for investigating strategies and trends for policy legitimation in education. The European Union-affiliated *Eurydice* network synthesises policy descriptions for the European countries in an online database that is widely used by policymakers. Analysing Eurydice data for assessment and grading policies, the paper discusses functional equivalence of grading policies and validity problems related to the comparison of such policy information. Illuminating the roles of the Swedish Government and a consultant in reviewing and recommending grading policies, the paper discusses new 'fast policy' modes of policy legitimation in which comparative data is used to effectuate assessment reform.

Keywords

Agency, collaboracy, consultancy, educational assessment, legitimacy, grading, policymaking, policy legitimation

Corresponding author:
Sverre Tveit, Department of Education, University of Agder, Postbox 422, Kristiansand, 4604, Norway.
Email: sverre.tveit@uia.no

Introduction

The Scandinavian countries have a long tradition of utilising expert and stakeholder committees that – in a collaborative fashion – review and propose policy changes, to legitimate governments' education reforms. After the turn of the millennium – and post the multiple 'PISA shocks' associated with international large scale assessments – a shift can be observed in Swedish policymaking towards more use of individual consultants and international agencies such as the Organisation for Economic Co-operation and Development (OECD) and the European Union. This article demonstrates how the traditional mode of policy legitimation – characteristic of Swedish policymaking – has been replaced by new *fast modes* of policy legitimation archetypical to global policymaking trends (Peck and Theodore, 2015).

A major and controversial issue for Swedish education in the last decades has been that of at what student age should schools embark on formal grading. When implementing new policies in this area it is particularly important for governments to ensure that the public and the teaching profession consider them as legitimate. Assessment and grading policies thus are suitable areas for investigating strategies and trends for policy legitimation in education. In this paper, we use assessment and grading policy as a case to show how comparative data on national states' education policies from, for example, the OECD or the European Commission, which we call *supranational agencies*, are used by policymakers to legitimate governments' ideologies.

The case of policy legitimation investigation created headlines in Swedish media in 2010–2011 as teachers and scholars opposed the government's proposed assessment reform (Dagens Nyheter, 2015; Lärarnas Tidning 2014). The Swedish Minister of Education at the time, Jan Björklund (the Liberal party), nominated the neuroscience professor Martin Ingvar from the prestigious hospital and medical school, the Karolinska Institute, as an expert to investigate potential implications of grading younger students. As a single expert, Professor Ingvar produced a green paper report (SOU, 2010) reviewing literature on this issue, backing the policy of embarking on formal grading in Year 6 (when students are 12 years old) instead of Year 8 (age 14), which was the policy at the time. Subsequently, in a memorandum in 2014 (Utbilningsdepartementet, 2014), Professor Ingvar examined students' ages when schools embark on formal grading in the European and OECD countries and recommended that Sweden further lower the use of grades to Year 4 (age 10). One of Ingvar's major arguments for this recommendation, also put forward by the government, was that countries performing better than Sweden in the Programme for International Student Assessment (PISA) had a system of 'early formal grading'. On the Swedish Television Broadcast's *Rapport* (equivalent to the *Six O'Clock News*), the Minister stated that:

> Almost the entire world grades their students earlier than Sweden does. Most countries grade from Year 1. Our neighbour country Finland grades students from Year 3 or 4. Countries that excel in PISA grade students very early. (SVT , Rapport [Swedish Televsion, Six o' clock news], 20 August 2014; authors' translation)

By nominating a distinguished neuroscience professor to review and recommend new grading policies – putting forward implicit causal claims that early formal grading leads to higher achievement – the Minister sought to effectuate an assessment reform.

Basing a controversial reform on implicit causal claims about the 'world situation' prompted researchers to investigate how the information about countries' grading policies was obtained. In the Ministry memorandum Professor Ingvar relied on implicit causal inferences when listing students' ages when schools embark on formal grading in the OECD countries. The reference given for that list was: 'see for example OECD 2013' (Utbilningsdepartementet, 2014: 37). However,

this OECD publication did not include such information. When called upon, Professor Ingvar said that the information came from the Ministry of Education, which had referred it to the OECD. When the Ministry was confronted with the lack of evidence in its references, the government official admitted that 'there unfortunately was a mistake in the reference list' (Lundahl, 7 October 2014, personal communication, U2014/5534/S School Unit, Ministry of Education; authors' translation). The government official then explained that the information was gathered partly from the Eurydice network and partly through contacts with the ministries of education in other countries. The question of how this information was constructed remained unanswered and the information therefore difficult to verify.

The aim of this paper is to explore the basis for the inferences drawn by the Swedish Ministry of Education and its consultant, and thus the legitimacy of the policy recommendations put forward with respect to reformed grading policy in Sweden. The case is illustrative of new trends with regard to national governments' policymaking and policy borrowing. We examine how policymakers legitimate change through the use of policy descriptions of other countries, as provided by supranational agencies, and the nomination of consultants to review and propose new policies based on this comparative policy data. The paper first elaborates on theoretical perspectives on policy borrowing and policy legitimation. Examples from educational assessment policymaking in Sweden and beyond are used to establish the distinctions between *collaboracy, agency* and *consultancy* modes of policy legitimation. Second, the data and methods used to analyse the contemporary case of policy legitimation of grading policy is outlined. Third, the paper undertakes a two-step analysis illuminating problems related to structure, labels and classifications of Eurydice data and the comparability of the meaning of 'grades' and 'grading' across countries. Fourth, the article discusses the emergence of new modes of policy legitimation in relation to other studies that observe similar developments within and beyond the Swedish context. The paper concludes by addressing the implications of these new modes of policy legitimation both for policymaking and research, and calls for educational research communities to give more attention to the (mis)use of Eurydice data in policy and research deliberations.

Theoretical perspectives on policy legitimation

Policy borrowing has received increased attention in educational research over the past few decades (Cowen and Kazamias, 2009; Schriewer, 2014; Steiner-Khamsi, 2004, 2010; Steiner-Khamsi and Waldow, 2012) in tandem with the growing international policy discourse sparked by international comparative studies of student achievement (Benveniste, 2002; Kamens, 2015; Petterson, 2008). While policy borrowing, strictly interpreted, refers to the situation when 'policy makers in one country seek to employ ideas taken from the experience of another country' (Phillips, 2004: 54), the term has digressed to a more general meaning related to how a nation's policy is influenced by other countries. One aspect of policy borrowing is that of legitimating national policy by referring to policies in other countries.

Jürgen Schriewer (1988) discusses how descriptions of foreign educational systems and their practices serve as frames of reference to specify appropriate reforms of a given nation's education policy. Simplifying Schriewer's (1988) perspectives, we can say that policies can gain or sustain legitimacy by referring to (1) scientific principles or to (2) values or value-based ideologies. With the latter, reaching a consensus can be difficult. Thus, externalisation to 'world situations' can be a useful strategy for objectifying value-based reasons for decision-making in education, accomplished in the forms of historical descriptions and/or statistical documentations that are recognised as scientific (cf. Schriewer, 1988: 62–72). As such, referring to other countries can make value-based policymaking more legitimate, as the values – through externalisation to world situations

– can *reappear* as scientific principles that have higher legitimation potential than values alone. For example, referring to Finland's grading system and to their high rank in PISA in order to legitimate a similar grading system in Sweden can be perceived as legitimation by externalisation.

Policy-borrowing literature often draws on neo-institutional theories that give attention to what DiMaggio and Powell (1983) described as different processes of *institutional isomorphism*. The theory of institutional isomorphism suggests that organisations (or countries) become similar due to external or hierarchical pressure (*coercive* isomorphism), through modelling other organisations (*mimic* isomorphism) or through organisational norms (*normative pressures*). In the case of national states' policymaking in education, these perspectives draw attention to how policymakers interact with one another, often facilitated by agencies such as the World Bank, UNESCO, the European Union and – especially these days – the OECD. These processes lead organisations to mimic one another's behaviour, or countries to borrow one another's policies.

Grek and Ozga (2010: 706) suggest that if one wants to predict and understand why and where policy is moving, one should be looking at the management of knowledge, rather than at policy itself. A nation's policy legitimation is often mediated by structural comparisons of which data reduction and classification may or may not be standardised (see also Lundahl, 2014). Grek (2013: 698) views comparison not simply as informative or reflective: 'In fact, it fabricates new realities and hence has become a mode of knowledge production in itself'. This type of policy legitimation is seldom made explicit, which becomes a problem in policy areas where the juridical and political terms are highly institutionalised and embedded in the nations' distinct traditions. Educational assessment, particularly the formal assessments and national instruments underpinning meritocratic procedures, often relies on 'taken for granted' information because 'all' members of the national political contexts have undertaken these assessments. This implicitness becomes particularly problematic when self-reported representations of national policies inform other countries' policymaking, which is the case with the Eurydice data that we investigate in this paper.

However, we believe that a strategic approach to synthesising and using other countries' policies, which Schriewer brings to our attention, is better labelled 'policy legitimation' than 'policy borrowing', acknowledging that the domestic setting – the need to legitimate a government's policies and ideologies – often is the important driver in this type of policy borrowing. We do not define policy legitimation as an entirely cynical and strategic part of policy deliberations. Even if politicians and other policymakers often have a most sincere belief that the outcomes of their reforms will be for the better, there are usually parallel strategies for maximising the chances that these beliefs will be received as legitimate.

Table 1 outlines the three modes of policy legitimation – collaboracy, agency and consultancy – as a framework for understanding different types and sources of legitimacy in national states' policymaking. The framework gives attention to the type of actors (A), their type of authority (B) and the type of institutional processes (isomorphism) (C) that produce the legitimacy. Further, Table 1 includes examples from Sweden's involvement in international research and policy deliberations and use of agencies and consultants (D), in addition to the identified emergences and peaks of the three modes of policy legitimation (E).

Collaboracy mode of policy legitimation

Collaboracy[1] is a mode of policy legitimation in which established actors – stakeholders, researchers and other experts – take the existing professional practice as the point of departure when reviewing policies and practices elsewhere. It can be understood through a Weberian perspective on traditional authority (Zymek, 2003). In Sweden there is a long-standing tradition of collaborating with stakeholders when the government formulates and reviews policies (Musial, 1999). Before

Table 1. Three modes of policy legitimation.

Modes	Collaboracy	Agency	Consultancy
(A) Actors	The government produces legitimacy by nominating stakeholders, researchers and other experts to review and recommend policy changes	The government produces legitimacy by cooperating with formal agencies which *fund, commission, synthesise, review* or *recommend* policy changes	The government produces legitimacy by nominating individual experts or private enterprises to review and recommend policy changes
(B) Type of authority	Representativeness and expertise as perceived by the public and professionals	Hierarchical (e.g. supranational towards national; national towards local)	Expertise in line with the targeted policy measures and promoted knowledge basis as defined by the government
(C) Institutional processes	Mimic isomorphism, not coercive. Inherent legitimacy maintained through tradition	Mimic isomorphism; sometimes object to coercive isomorphism	Selective, mimic isomorphism (selective modelling)
(D) Examples from Sweden	SOU, 1977: 9	Eurydice, 2014a OECD, 2015 OECD, 2013	McKinsey & Company (Barber and Mourshed, 2007) SOU, 2010: 52; Utbilningsdepartementet, 2014
(E) Emergences and peaks in Sweden	The 20th century → (The assessment and testing reforms in the 1960s and 1970s)	2000 → (The PISA shocks and subsequent 2011 reform)	2010 → (The 2011 educational reform and further change of grading age in 2014)

OECD: Organisation for Economic Co-operation and Development; PISA: Programme for International Student Assessment

the government draws up a legislative proposal for a new policy, it may choose to appoint a special expert or group – officially known as a one-man committee of inquiry or a commission of inquiry – to investigate the issues in question. Reporting on matters in accordance with a set of instructions laid down by the government, these operate independently and may include or co-opt experts, public officials and politicians. The reports are published in the Swedish Government Official Reports series (Statens Offentliga Utredningar (SOU)). After a committee has submitted its report to the responsible minister, it is sent to relevant authorities, stakeholders and the public for consideration. These are given an opportunity to express their views before the government formulates and presents a legislative proposal to parliament. As such, reforms undertaken are prepared by everyone who is part of the education system. Characteristic of this mode of policy legitimation is when an expert committee in the 1970s reviewed and discussed grading age for up to a decade before conclusions and policy recommendations were presented to the government (Lundahl, 2006).

Agency mode of policy legitimation

The *agency* mode of policy legitimation involves formal agencies that shape policymaking by funding or commissioning policy interventions, by synthesising policy data from different countries, or by reviewing and recommending policies. In the context of national states' policymaking

we draw on Dale's (2005) use of the concept *supranational* to give attention to a range of agencies which have been observed to increase influence on national states' policymaking (Grek, 2009, 2013; Ozga et al., 2011), such as the OECD, the European Union, the World Bank and UNESCO. Whether these international features of policy brokering take the form of Europeanisation (Grek and Lawn, 2012) or globalisation (Dale, 2005), they both reflect and condition national states' governing processes.

In the agency mode of policy legitimation, mimic isomorphism may be amplified by coercive power of the respective supranational agency, for example, through conditional benefits it offers. An example of coercive power is when the World Bank requires education systems to transform themselves to meet the demands of the global knowledge economy (Robertson, 2005) and western neo-liberal fiscal policies (Jones, 2004). This type of the agency mode of policy legitimation may effectively enforce countries to accept supranational agencies' testing and accountability policies to receive benefits from, for example, the World Bank (Benveniste, 2002).

In this paper, we are, however, mainly concerned with mimic isomorphism, both with respect to the construction of comparative policy data that form the basis for modelling and with the way such policy information is used. This type of legitimation has become more prominent in the new millennium. Cussó and D'Amico (2005) observe increased competition between such agencies, which led UNESCO to align with the other main knowledge brokers and collectors of educational statistics. Dale (2005: 119) notes, 'This places great power in the hands of the agencies setting up the statistical variables that would determine what the "proper" outcomes of education should be'. 'Quick' global and European-level policy comparisons increasingly inform national states' policymaking (Grek and Lawn, 2009, 2012; Rizvi and Lingard, 2010). Lundahl and Waldow (2009) identify how 'quick languages', for example, in comparison with national states' outcomes on international tests, frame and make educational policy discourse accessible to wider circles of participants. While public attention seldom reaches below the surface of these outcomes comparisons, government officials and politicians delve deeper into the datasets in search of recipes for successful policies. Thus, international agencies, such as the European Union and the OECD, are increasingly used by policymakers to provide synthesised comparative data that can be used in national reform agendas.

While this mode can be associated with the large influence of the PISA study (Grek, 2009), the OECD's role in national states' agency mode of policy legitimation, particularly in relation to accountability policies, long preceded the PISA tests. Established in 1968, the OECD's Centre for Educational Research and Innovation began providing policy recommendations to member countries (Lundgren, 2011). Yet, following the implementation of the PISA tests, the countries increasingly gave emphasis to the agency's policy reviews and recommendations (Pettersson, 2008). Drawing on Börzel and Panke (2013), Prøitz (2015) demonstrates a sequential approach of uploading and downloading that shaped the OECD's (2013) policy review *Synergies for Better Learning*. In Sweden, the OECD report *Improving schools in Sweden: An OECD perspective* (OECD, 2015) is a recent example of how the government uses the OECD to review and recommend policies to legitimate its policies. The European Union-affiliated Eurydice network does not have an active role in recommending policies; however, its synthesised policy descriptions of the European countries are located in an online database that is widely used by policymakers to inform decision-making and by researchers undertaking comparative studies.

Consultancy mode of policy legitimation

The *consultancy*[2] mode of policy legitimation is related to governments' utilisation of individual consultants or private enterprises to review and recommend policies. Lindblad et al. (2015) observe

that global private enterprises such as McKinsey & Company have become increasingly involved in policymaking in recent years. Gunter et al. (2014: 519) observe that consultants are increasingly recognised as 'external knowledge actors who trade knowledge, expertise and experience, and through consultancy as a relational transfer process they impact on structures, systems and organisational goals'. Coining the term 'consultocracy', Hood and Jackson (1991) identified a trend in which 'non-elected consultants are replacing political debate conducted by publicly accountable politicians' (Gunter et al., 2014: 519). Reviewing public policy studies, Gunter et al. (2014) observe that rapid, radical and often incoherent changes in public administration can be understood in view of consultancy businesses playing a substantial role in both responding to and generating reform. They argue that this has become the new 'normal' context of policymaking in the United Kingdom. We call these developments a new *consultancy* mode of policy legitimation.

McKinsey & Company's report *How the world's best performing school systems come out on top* (Barber and Mourshed, 2007) is an example of the increased influence of consultancy enterprises that also received extensive public attention in the Swedish media[3]. As such, policymaking in Sweden is increasingly conditioned by 'the public eye' (Rönnberg et al., 2013: 178). The need to accommodate the press and social media's demands for brief information is characteristic of the consultancy mode of policy legitimation.

While in the UK it has become more common to nominate private enterprises to legitimate policy changes, in Sweden the nomination of one-man inquiries is typical of what we call consultancy[4]. The nomination of Processor Ingvar to undertake a review entitled *Biological Factors and Gender Differences in School Outcomes* (SOU, 2010: 52) and produce the memorandum *A Better School Start for All: Assessment and Grading for Progression in Learning* (Utbilningsdepartementet, 2014) – reviewing and recommending new policies for grading age – are recent examples of one-man inquiries in the field of educational assessment and grading.

The proposed classification of three modes of policy legitimation can be helpful in coming to terms with different strategies for undertaking and legitimating policy changes. In historical studies, such as the above brief examples from Sweden, the modes can be used to identify eras and milestones. They enable us to pinpoint how the *time* allowed for policy deliberations in the *collaboracy era* marks a contrast to the contemporary 'fast policy' era, where approaches that we have called *agency* and *consultancy* are more efficient modes of policy legitimation. The concepts contribute to the well-established literature on policy borrowing between countries. However, the modes we identify are principal ones that can also relate to domestic and local settings. Furthermore, the modes of policy legitimation should be understood as typologies that can occur to various extents, independently and simultaneously. For example, the collaboracy feature of circulating policy recommendations (that are proposed by expert and stakeholder committees) on referral still operates in Sweden, although with less legitimating power due to the new modes of policy legitimation that confront the traditional values and approaches to policymaking that the profession was accustomed to. In the following sections, the focus is on how consultancy operated in tandem with agency when the Swedish government used comparative data on countries' assessment policies to effectuate an assessment reform.

Data and methods

The information provided by the Eurydice network is the main basis for the empirical analyses in this paper. As we have shown, data from Eurydice was used in attempts to effectuate an assessment reform implementing formal grading in lower school years by referring to 'world situations'. To scrutinise the validity of implicit causal claims that this policy change relied upon, we undertake a two-step empirical investigation guided by two research questions. First, how is the representation

of countries' policies conditioned by the Eurydice database's headings and classifications? Second, when asked to describe the system of formal grading in a country in Eurydice, what kinds of descriptions do the various countries provide?

Facilitated by the European Commission's Education, Audiovisual, and Culture Executive Agency, Eurydice provides European-level analyses and facilitates comparison of education policies in Europe developed to assist policymakers responsible for national education policies (Eurydice, 2014a). Eurydice can be described as a form of Web-based encyclopaedia using a structure similar to, for example, the country systems report in the *International Encyclopaedia of Education* (Husén and Postletwaite, 1985, 1994; see also Lundahl, 2014).

Research into social knowledge has often been concerned with the (micro) processes that shape scientific knowledge (e.g. Camic et al., 2011). Encyclopaedias are often claimed to be collections of facts – that is, a knowledge storeroom. Typically, we perceive facts as 'unconstructed by anyone' (Latour and Woolgar, 1979/1986). But producing an encyclopaedia is not a straightforward and simple editorial process. Sections, headings, topics and the structure of the thematic articles are constantly changed based on new insights and on circumstances beyond anyone's control. A better way to frame the knowledge in an encyclopaedia would be to understand it as a product of a specific epistemic culture – the actual and theoretical conditions of the production of knowledge (Knorr Cetina, 1999). To put it differently, it is not only 'truth criteria' (or the preservation/development of knowledge) that can be seen as a reason to produce an encyclopaedia. Encyclopaedias can be treated as we treat other kinds of knowledge. Knowledge is geographical, sociological and chronological (Burke, 2012). In other words, we can expect editors of an encyclopaedia such as Eurydice to struggle with geographical and periodical frames, translations and issues in deciding on relevance and limitations of content and of contributors.

There is not much written about the use of this type of comparative data describing countries' educational systems (Lundahl, 2014). To investigate the quality of the comparative data used to effectuate the assessment reform in Sweden, we investigated the Eurydice data sources that the Ministry and its consultant used, and established an overview of European countries' policies on educational assessment and grading. In addition to the respective country sections of the Eurydice material obtained during fall 2014, our analysis draws on two Eurydice schematic diagrams displaying the 'Compulsory education in Europe' (Eurydice, 2014b) and the 'Structure of the European education systems 2014/15', published in November 2014 (Eurydice, 2014c). We classified and structured each country's information on primary and secondary education to provide a comparable overview, as shown in Appendix 1. The initial structure comparison of the Eurydice content demonstrated vast differences between the countries' frameworks of reporting. This implied a need to undertake the analyses through an abductive approach to establish suitable categories for classification (Schreier, 2012).

Schriewer (1988: 33–34) emphasises that comparative research should 'not consist in relating observable facts but in relating relationships or even patterns of relationship to each other'. Accordingly, procedures were undertaken for establishing the relationship between different segments of policies (e.g. school structure and tracked programmes), aiming to shed light on the (lack of) functional equivalence of the compared constructs (Schriewer, 2003). We do not possess the capacity and access to obtain insight into all policy structures and legal and political terminology, and thus we cannot give full accounts of these relationships. However, the vast differences that can be identified are sufficient to substantiate validity problems related to the comparability of education systems and the policy descriptions characterising these.

Substantial validity problems are related to the construction and use of data intended to facilitate such comparisons, including methodological challenges with regard to the classification of information from national education systems within standardised categories. Thus, different

interpretations across the countries may cause variations in reporting on, for example, student age when schools embark on formal grading or the degree of student retention. These problems of constructing and using comparative policy data are equally complex for us as researchers and for policymakers aiming to legitimate policies by referring to other countries. Rather than aiming to achieve clear-cut distinctions that warrant generalisations, which may be what policymakers desire, our goal is to illuminate the immense (and implausible) challenge of painting a flawless portrait based on this mosaic of information.

Analysis of the construction and use of comparative policy data

In this section, we first analyse how the qualitative information about grading policies is structured in Eurydice and illuminate problems related to the construction and use of this 'encyclopaedia' with respect to the structure of headings and content and its implications for the comparison of countries' policies. Acknowledging these shortcomings, we continue comparing Eurydice's qualitative information about countries' policies for student age at which countries embark on formal grading.

Analysis I: structure, labels and classifications in the Eurydice encyclopaedia

Appendix 1 provides the full accounts of the analysed Eurydice data. The methodological challenges of using Eurydice include but are not limited to the problems of classifying and synthesising information. Compiling and comparing information on the grading systems in Europe is affected by differences in: the age at which students start school, the length of compulsory education and its structure (e.g. comprehensive versus tracked programmes), associated selection procedures, practices for issuing formal certificates in early years, and the status and use of these. Eurydice reports, such as the comparison in 'National testing of pupils in Europe' (Eurydice, 2009), use other sources to substantiate inferences. In this study, however, we have purposively restricted the investigation to qualitative data in the online Eurydice database to illuminate problems of using (primarily) this type of data as the basis for comparison and borrowing.

Thus, a first challenge was to determine the age at which children start school, as this is regulated in different ways across the countries. In many countries preschool is compulsory; however, preschool is integrated in the compulsory education system to various extents. The length of compulsory education can therefore vary substantially depending upon how one classifies starting age. This, in turn, has implications for the determination of when schools embark on formal grading in the various countries.

In Eurydice, countries are classified according to education structure in order to facilitate comparison across countries. Sweden is classified as 'single structure', which relies on the assumption of primary and lower secondary being integrated, in contrast to countries such as England and Germany, which are classified and distinguished as 'primary education' and 'secondary education'. This reflects a fundamental structural difference between the education systems, where a vast number of countries, like Sweden, have a comprehensive school system, whereas others, such as Germany and England, have a clearer separation at the intersection between primary and secondary education.

These differences are important to consider when comparing countries' grading policies. In education systems where grades inform the admission to further education, grades given in Year 4, 5, 6 or 7 may have a different role from those in the Nordic countries, where there is no such transfer within compulsory education. In countries that employ a firm separation between primary and secondary education, the grades may serve an important role in admitting students to differentiated

secondary programmes. These purposes of grading do not exist in countries with an integrated compulsory education system (called single-structure education in the Eurydice data).

However, among countries that in Eurydice are labelled and classified as the same type of education system structure, there are substantial differences. To better comprehend the structural differences that in Eurydice are distinguished in two categories (single structure and primary/secondary), we introduce a third category to distinguish between non-single structure countries that have a firm separation between primary and secondary education and countries that also track students when they commence (lower) secondary education. We use the following classification for countries' education structures:

1. Single structure (comprehensive education)
2. Primary secondary structure (comprehensive lower secondary programmes)
3. Tracked secondary structure (differentiated lower secondary programmes)

In the first group, we find the countries Bulgaria, Croatia, Denmark, Estonia, Finland, Hungary, Iceland, Norway, Portugal, Slovakia, Slovenia, Spain, Sweden and Turkey. In the primary secondary group, we find Cyprus, France, Greece, Lithuania, Malta, Poland, Romania and countries in the United Kingdom. In the tracked secondary group we find the countries Austria, Belgium, Germany, Ireland, Lichtenstein, Luxembourg and The Netherlands.

When looking further into some countries it becomes evident that the Eurydice data may be misleading in its classification. One example is Latvia, which we first classified as a single-structure country based on the information from Eurydice. When investigating further, however[5], it appears that students are differentiated already at lower secondary and thus should be classified into group 3. Italy is another example. It is not classified by Eurydice as a single-structure education system; however, according to the Eurydice data it is possible for schools to provide integrated comprehensive schooling[6]. Thus, for these two countries we have used the label 'inconsistency'. This label may prove to be appropriate for several other countries too; Latvia and Italy are just mentioned as examples, to explicate the problems of classifying countries based on the information structure of the Eurydice database.

With respect to students' age when they are differentiated (see Appendix 1), we generated the data from What was reported in the Eurydice section called 'progression of pupils'. While we could classify most countries based on this information, nine countries had not provided the necessary information. Furthermore, even within the primary secondary structure countries we find contradictory information in the Eurydice data. For instance, France has a complex provision of different types of schools within the public education system. This is, however, not provided based on classical 'tracked' differentiation (such as in Germany). Thus, it was classified as primary secondary, but with differentiation at age 11 (after six years in school), hence a type of hybrid between the Primary Secondary and Tracked Secondary structures. England was even classified as Primary Secondary, with differentiation at age 16 (after 12 years of schooling), as Eurydice does not report that English students undertake secondary schooling that leads into different tracks. Looking closer into this, however, the provision of private (Academy) schools implies a substantial segregation of students based on socio-economic premises that in effect leads to earlier differentiation for many students.

The above examples are mentioned to bring attention to the complexity associated with this type of data, and the need to use it with caution. Comparing a basic item such as the age at which children start school is a complex issue to which Eurydice and the OECD give much attention as it has substantial implications for comparing national states' educational achievements. This has implications for the comparison of what year countries begin formal grading of students. The examples

Table 2. Student age at commencement of formal grading[7].

Year	Age	Countries
0	**5**	One country: Northern Ireland*
1	**6**	Nine countries: Cyprus, England, France, Hungary, Italy, Poland, Romania, Wales, Austria
2	**7**	Four countries: Belgium (French), Finland**, Luxemburg, Turkey, Germany (majority of the states)
3	**8**	Four countries: Greece, Latvia*, Malta, Slovenia
4	**9**	Two countries: Portugal, Slovakia
5	**10**	
6	**11**	Two countries Lichtenstein, Lithuania
7	**12**	One country: Sweden
8	**13**	Two countries: Norway, Finland**
9	**14**	One country: Denmark

*Northern Ireland and Latvia are classified as starting in Year 0 to standardise the classification of student age in respective school years.
**Finnish school children commence schooling at the age of seven years. In Finland, school policies for embarking on formal grading vary locally. It is optional to give formal grades in Years 1–7 but compulsory to do so from Year 8.

illustrate just a few of these complex factors that condition the premises of grading policies. They illuminate how difficult – or outright impossible – it is to arrive at a comparable notion of 'grading age'. One should therefore use the information provided in Appendix 1 and Table 2 with utmost caution, and note that they have been generated *to substantiate variations* rather than to facilitate comparison.

Analysis II: grades as functional equivalent constructs of comparison?

When we examine the grading policy information in Eurydice, what on the surface appears as uncomplicated information – which the Swedish Ministry of Education and their consultant used as hard facts about grading – has embedded several nuances. Grading usually refers to a formal assessment at the end of a semester or school year based on the student's achievement in relation to specific criteria or ranked relative to other students; for example, Brookhart (2015) defines a grade as given by teachers representing the sum of many achievements and not just a single test score. She perceives it as an internal tool of assessment rather than an external one, such as for various national tests and examinations. Other scholars have other perceptions of what a 'grade' is. So too do policymakers.

Thus, 'grading' is an extremely complex phenomenon that takes many different forms. For example, a policymaker may define it as formal grading when a student sits through a national test in, for example, Year 2, whereas other policymakers, or a person in charge of synthesising the information, may view the national test as a rare exception from the normal environment in which formal grading does not occur.

A problem, which indeed is what the Eurydice network aims to help policymakers with, is the issue of sharing information that for the most part is available in the nations' domestic languages. For researchers, it is important to examine other sources, particularly scholarly articles about countries' education policies, to control for misinterpretation or biased representations of national policies (Dale and Robertson, 2009). Thus, mastering the relevant language is pivotal. The problem of language and translations causes substantial 'noise' to the 'uploading' of policy data in Eurydice. In some cases, the English used is downright poor. Thus, it is with the greatest care that we may

draw some conclusions based on comparisons of countries' grading systems. This applies to both educational research and policymaking.

When studying how the European countries describe their grading systems in Eurydice, we first note that 11 countries do not even mention at which point they start grading student achievements. This means that it is difficult to control general statements about the number of countries that give grades, or about when and how such grades are given. This information is sometimes only available through hyperlinks to the countries' national education authorities, in their respective native languages. Thus, nearly one-third of the countries do not appear to view grading age as important information to be provided in the Eurydice database.

Nevertheless, as shown in Table 2, it is clear – and in line with the Swedish Ministry and its consultant's claims – that most countries report student grading commencing earlier than in Sweden. This, however, is for different reasons and done in very different ways. There are various systems and principles for grading students (see also Lundahl et al., 2015). Below, we use Eurydice data to describe a range of countries that embark on formal grading in the early years, giving attention to the vastly differing descriptions hiding beneath these numerical comparisons.

Austria provides one conclusive grade at the end of Year 1 (age six years), sometimes together with oral supplements. It is not until Year 2 that the children receive grades for all subjects. *Latvia* also employs a system in which the students receive a more qualitative overall grade aged five and then receive grades on a scale ranging from 1 to 10 from Year 1 (age six) in the native language and mathematics. For the other subjects, the teachers give more qualitative judgements up until Year 3 (age eight), when schools embark on formal grading across all subjects.

In *Cyprus*, the students receive a progress certificate each year from Year 1 (age six), which shows whether the student passed the Year or not. Without this, they are not allowed to move up to the next class. At the end of Year 6 (age 11), the students receive a leaving certificate from primary school. In *Turkey*, students take a proficiency test at the end of Year 1 (age six) to determine what education they will receive in Year 2 (age seven). At the end of each year from Year 2 to Year 8, students take an examination in which they must receive the judgement Fair (3) to be allowed to complete the year. Also, *Lithuania*, although not giving grades in early years, has approved results of Year 1 as a prerequisite for moving up. Students can move up before completing Years 1–3 if they are believed to be able to cope with what the curriculum covers in these years.

In *Finland*, students have the right to a judgement from their first Year (age seven). The forms of these judgements are determined locally, and it is up to the local authorities to decide whether they want to give grades or oral assessments until Year 7. From Year 8 (age 13) onwards, grades are always numerical. *France* employs a system were the students' results on varies tests and examinations are summarised in a book (*livret scolaire*). This book is used as part of communications that teachers have with the children's parents to give a continuous account of the child's development. *Hungary* has a system in which teachers must give students regular 'marks' during the year and summarise them at the end with a rating. In *Italy*, at the end of each study period the students receive a summarised assessment document. Students are also graded on their conduct. In *Poland*, students are graded on a scale of 1 to 6, starting in Year 1 (age six).

These examples demonstrate that what, explicitly or implicitly, are perceived as grades are not functional equivalents when policymakers upload and download information to the Eurydice database and when policymakers and researchers use this information to compare countries' grading policies. The different meaning of grading can be further illustrated with the examples of Cyprus, France and Sweden. In Cyprus grading is connected to retention, whereas in France it is considered a more formal feedback to the parents. In Sweden, the latter function, for example, is already achieved through what is called 'written judgements' (skriftliga omdömen): qualitative reports that

parents receive from Year 1; but Swedish policymakers and researchers would not call that 'grading' in their Eurydice report.

We can draw the preliminary conclusion that the various content items related to grading policy are not functionally equivalent (Schriewer, 2003b), which undermines the validity of brief policy comparisons based on this type of data. Policymakers and researchers should consider that there are great differences with respect to the use of grades, when comparing 'grading age'.

Discussion: the emergence of new modes of policy legitimation

The aim of this article was to demonstrate new trends with regard to national governments' policymaking and policy borrowing, with increased use of individual consultants and international agencies to legitimate policy changes. As demonstrated above it is unlikely that a valid comparison of countries' 'grading age' was even comprehensible for the Swedish policymakers. To make such a valid comparison would require substantial resources with respect to information generation, translation and validation procedures, requiring time that would far exceed what was assigned for the investigation.

Despite our illumination of the lack of a valid premise for drawing inferences from Eurydice with regard to student age at commencement of formal grading, we have also undertaken a correlation analysis relating the generated Eurydice information to the countries' PISA scores (Lundahl et al., 2015). We find no evidence for the Minister's remark that 'early grading countries' excel in PISA or that there is a causal relationship between these factors (which even in the event of any covariance would be a naïve and speculative inference to draw). Our analyses suggest that even if it is possible to rely on Eurydice data to reveal some fundamental differences when it comes to grading systems in Europe, there are large variations in how the countries describe these policies. The Eurydice data simply do not provide essential information on the reality of grading policies in Europe.

From the above theoretical and empirical accounts we can observe a historical development and shift in the modes of policy legitimation used by Swedish policymakers. The strong Swedish tradition of using stakeholder and expert committees that undertake comprehensive investigations declined after the beginning of this millennium. In the 1970s, the issue of formal grading was thoroughly discussed and investigated in Sweden in a traditional collaboracy fashion. Stakeholders and experts would work side by side in government committees for almost a decade to investigate ideological disputed questions related to student age at commencement of formal grading. Fast-forward to the post-millennium era of policymaking, and we see that both agency and consultancy modes of policy legitimation, within half-year sequences, are at play when seeking to effectuate reforms. Increasingly, governments have nominated experts rather than stakeholders to the committees. Nowadays, three-quarters of the expert inquiries (SOUs) are one-man inquiries conducted by single experts in one- to one-and-a-half-year sequences (Petersson, 2013). This has been particularly evident when it comes to the field of educational assessment and grading, wherein each of the four major government reports over the past decades has been formulated at a steadily increasing pace, ranging from the nine-year inquiry of the 1970s (SOU, 1977: 9) to the one-year or less inquiries of today (SOU, 2010: 52; Utbilningsdepartementet, 2014). Correspondingly, these have changed from being parliamentary to one-man investigations (Lundahl and Jönsson, 2010).

The increased emphasis on *agency* mode of policy legitimation may also indicate a shift from what Waldow (2009) describes as a Swedish distinct tradition of 'silent borrowing'. Comparing committee reports (SOUs) in the 1960s and 1970s, Waldow (2009: 486) argues that 'Swedish political culture in the second half of the twentieth century was characterised by the belief in the

rational, more or less non-ideological steering of economic, social and educational policy'. The recent OECD (2015) report and the consultant's review of OECD data may indicate a change toward more use of OECD data in SOU policy deliberations. On a more general level Ringarp and Waldow (2016) noted that, before 2007, international reference points were almost never used as an argument for reform in Swedish policy-making, despite Sweden was participating in many ways in the international education policy-making mainstream. This seems to have changed around the year 2007, when the 'international argument' became prominent in the education policy-making discourse as a legitimatory device and justification for change.

As government-nominated committees (SOUs) are long-standing institutions in Swedish policy deliberations, there is much power associated with these committees. The changed composure of these committees, however, may not be fully acknowledged among the stakeholders and public. Green papers signify an authority and legitimacy that may be more reliant on profound tradition than reflected in the contemporary procedures for nominating committee members. Thus, governments can nominate ideological allies as experts with a mandate to produce desired policy recommendations based on 'fast policy' reviews provided by, or based on data obtained from, agencies such as the OECD, the European Union and associated networks (such as Eurydice). Mediated through the recognised SOU institution and format, these policy reviews and recommendations can provide scientific legitimacy for policy changes in line with the government's ideology. It is a 'perfect medium' for 'levelling up' values and ideologies – through 'world situations' – to scientific evidence that in turn can inform and legitimate reforms.

Ringarp and Waldow (2016) argue that in countries like Sweden, with self-confidence as a pioneer country in education, facing PISA scores below average undermines this self-confidence and consequently makes externalising to world situations more attractive as a legitimatory resource. Large-scale assessments may appear particularly attractive as a legitimatory reference, they write, 'as referring to them combines externalising to world situations on the one hand with externalising to scientificalness on the other. Thereby, two powerful sources of legitimacy are tapped at the same time' (2016: 6).

As the PISA studies undergo rigorous validation procedures with respect to both quantitative and qualitative comparisons, these studies have shaped an era of general 'trust in numbers' (Lundahl and Waldow, 2009). This gives reform arguments more legitimacy when based on supranational agencies such as the OECD and the European Union (more or less) irrespective of whether data undergo rigorous validation procedures. Given the national states' extensive use of such comparisons in policymaking, there are reasons to critically examine the thoroughness of 'express' reviews and recommendations. It appears that simply referring to supranational agencies such as the European Union and the OECD as the basis for this type of knowledge brokering provides the desired legitimacy despite being based on vague, incomplete and at times misleading information. Our analysis demonstrates that comparisons of qualitative policy descriptions in Eurydice should be treated with utmost caution. To have some comparative value, the countries' information on at what student age schools embark on formal grading should be tied explicitly to International Standard Classification of Education levels (as done for standardised tests in the Eurydice 2009 report) and the meaning of grading explicitly defined.

The above analysis also sheds light on substantial problems related to conceptualisation and classification when constructing policy information. This further relates to how agency and consultancy modes of policy legitimation can operate when governments use comparative policy data to show that reform ideas are in line with either the normal or the successful 'world situation'. Our empirical analyses demonstrate that this information is difficult to validate, and we claim that in some cases the objects of comparison – such as student age when schools embark on formal grading – are not functionally equivalent at all. Lundahl et al. (2015) discovered in a systematic research

review that there is limited comparative research on assessments in general, and on grading in particular – a situation that also cements the problem of implicit borrowing; an implicitness that becomes problematic given that countries' 'lending' (Steiner-Khamsi, 2010) or 'uploading' (Prøitz, 2015) of national policies inform other countries' policymaking.

Unfortunately, Eurydice does not call our attention to the principal challenges associated with the quality of the qualitative data that it provides to policymakers and researchers. Eurydice's website slogan, *Better knowledge for better policy*, creates high expectations of sound policy data (Eurydice, 2015). While this may be true for the quantitative information provided, and its synthesised reports, our investigation suggests that the qualitative descriptions of the countries' educational assessment policies do not meet these expectations. We argue that Eurydice fails to meet common academic standards of transparency with regard to the construction of comparative policy data, and furthermore that it should have been explicit about potential threats to the validity of the comparisons and information provided. This becomes particularly problematic when this type of comparative policy data – synthesised and facilitated by supranational agencies – is used by consultants working with limited resources, with short deadlines and a mandate with a narrow focus, in order to shed light on political issues on behalf of governments.

Gunter et al. (2014) identify the relationship between the state, public policy and knowledge construction as an important site for analysing the role of consultants. The present study has unravelled an example of the interplay between these actors and processes. It substantiates a phenomenon Grek (2013) has identified in which previous accounts of 'knowledge and policy' or 'knowledge in policy' shift to a new reality where knowledge *is* policy. 'It becomes policy, since expertise and the selling of undisputed, universal policy solutions drift into one single entity and function' (2013: 707). We have proposed three concepts of policy legitimation strategies that help come to terms with the interplay between national policymakers, their use of comparative data provided by supranational agency actors such as the European Union and the OECD, and the use of consultants to review and utilise this information in policy recommendations. The collaboracy, agency and consultancy modes of policy legitimation can moreover be viewed as characteristics and milestones related to how the act of policy legitimation has emerged over time.

Conclusion

Our theoretical discussion highlighted the improbability that comparative policy data translate well across countries, due to the distinct different national contexts of formulation and interpretation. Our empirical investigations illuminated validity problems related to structuring, labelling and classification of policy information. Further, we have demonstrated significant problems of identifying conceptual equivalence with respect to the meaning of 'grading'. In other words, the validity in the comparisons of different nations' grading systems is low. Thus we have illuminated that the Swedish policymakers' approach to effectuate the assessment reform – claiming a need for coherence with the European and global 'normality', as suggested by a one-man expert review and based on Eurydice data – cannot be substantiated with valid scientific evidence.

In sum, the paper offers a remarkable example of how agency and consultancy modes of policy legitimation can operate in tandem when policymakers utilise comparative data to effectuate reforms. We have demonstrated that this approach relied on misinterpretations and inferences based on global and European-level information concerning policy structures that do not stand the test of scrutiny. As such, we have illuminated an example of European policy isomorphism (DiMaggio and Powell, 1983) that relies on, at best, wrong inferences that nevertheless created new national semantics for governing education that was used to effectuate a controversial assessment reform.

We have unravelled fundamental problems of transparency associated with this type of policy information with implications far beyond the Swedish case of policy legitimation investigated in this paper. What then causes governments to use this type of policy information? In a 'fast policy' era in which 'the complex folding of policy lessons derive from one place into reformed and transformed arrangements elsewhere' (Peck and Theodore, 2015: 3) – at an all-time high speed – governments are constantly searching for ways to legitimate their reforms. At the same time, the influence from supranational agencies such as the OECD and the European Union is increasing, the market for education consultancy companies is growing and, furthermore, individual consultants are more commonly used for inquiries nowadays. Single experts serve as governments' consultants, with limited time frames, and risk reproducing data and information of poor validity due to shallow contextual understanding.

It can be questioned whether national policymakers, supranational agencies, consultancy companies and independent consultants have the time and capacity to fully grasp the fundamentally different premises of education systems that exist across countries. It is imperative that policymakers ask these types of questions of agencies and consultants to validate the policy reviews and recommendations they commission and receive. It is important for policy researchers to acknowledge that this 'fast policy' era may pose a threat to the legitimacy of comparative reviews and the scholarship of comparative educational research itself. Thus, we argue that the construction and use of comparative policy data in European policymaking should be given higher priority and be examined thoroughly in future research by the comparative education community.

Acknowledgements

The authors are indebted to members of the research group Curriculum Studies, Leadership and Educational Governance (CLEG) at the University of Oslo – in particular, Adjunct Professor II Helen M Gunter – and members of the research group Education and Democracy, Örebro University, and to doctoral student Judit Novak at Uppsala University, for comments on the manuscript. The paper is part of the PhD project Assessment and Selection in the Scandinavian Education Systems (ASSESS), funded by the University of Oslo.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work was supported by the Swedish Research Council (grant number [dnr/ref] 2014-1952, From Paris to PISA. Governing Education by Comparison 1867-2015).

Notes

1. *Collaboracy* is a twist on the term *collaborative*, constructed to operate in tandem with the *agency* and *consultancy* modes of policy legitimation. It may well be described as *collaborative policy legitimation* in other contexts.
2. The concept is similar to that of *Knowledge Brokers* (Hargadon, 1998), but whereas these denote a mutual relationship between a 'buyer and seller' of knowledge of any kind, consultancy rather implies 'expert advice' nominated by an authority with a certain agenda.
3. See for example: http://www.skolledarna.se/Skolledaren/Artikelarkiv/2011/Ny-rapport-av-McKinsey/

4. Other examples are Leif Davidsson's review of the national curriculum (SOU, 2007: 28), Anita Ferm's review of general upper secondary education (U, 2007: 1) and Metta Fjelkner's review on order and conduct (U, 2014a: B).
5. 'The last years of basic education (grades 7–9) can also be acquired in *ģimnāzija*, which principally offers three years of full-time general upper-secondary education to students aged 16 to 19 […] It is possible to obtain compulsory education also in vocational schools.' https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Latvia:Single_Structure_Education_(Integrated_Primary_and_Lower_Secondary_Education)
6. 'Comprehensive institutes can be set up with the aim of ensuring didactic continuity within the same education cycle, consisting of a primary school, a lower secondary school and a pre-primary school, all run by a single school manager': https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Italy:Organisation_of_General_Lower_Secondary_Education
7. Not accounted for in Eurydice in 2014: (11 countries): Belgium (Flemish), Belgium (German), Bulgaria, Czech Republic, Croatia, Estonia, Ireland, Iceland, The Netherlands, Scotland, Spain. Note that the Eurydice database is updated regularly, thus the countries' representation may have changed substantially.

References

Barber M and Mourshed M (2007) *How the World's Best-Performing School Systems Come Out On Top*. [Location unknown]: McKinsey & Company.

Benveniste LA (2002) The political structuration of assessment: Negotiating state power and legitimacy. *Comparative Education Review* 46(1): 89–118.

Börzel TA and Panke D (2013) Europeanisation. In: Cini M and Borragán NP-S (eds) *European Union Politics*. Oxford, UK: Oxford University Press, pp.115–127.

Brookhart SM (2015) Graded achievement, tested achievement, and validity. *Educational Assessment* 20(4): 268–296.

Burke P (2012) *A Social History of Knowledge II: From the Encyclopédie to Wikipedia*. Cambridge, UK: Polity.

Camic C, Gross N and Lamont M (2011) *Social Knowledge in the Making*. Chicago: University of Chicago Press.

Cowen R and Kazamias AM (2009) *International Handbook of Comparative Education*. London: Springer.

Cussó R and D'Amico S (2005) From development comparatism to globalization comparativism: Towards more normative international education statistics. *Comparative Education* 41(2): 199–216.

Dagens Nyheter (2015) Inga hållbara argument för betyg redan i fjärde klass [No valid arguments for grading already in Year 4]. Available at: http://www.dn.se/debatt/inga-hallbara-argument-for-betyg-redan-i-fjarde-klass/

Dale R (2005) Globalisation, knowledge economy and comparative education. *Comparative Education* 41(2): 117–149.

Dale R and Robertson S (2009) Beyond methodological 'isms' in comparative education in an era of globalisation. In: Cowen R and Kazamias AM (eds) *International Handbook of Comparative Education*. Dordrecht: Springer Netherlands, pp.1113–1127.

DiMaggio PJ and Powell WW (1983) The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields." *American Sociological Review* 48(2): 147–160.

Eurydice (2009) National testing of pupils in Europe: Objectives, organisation and use of results. Available at: http://eacea.ec.europa.eu/education/Eurydice/documents/thematic_reports/109EN.pdf (accessed 9 November 2015).

Eurydice (2014a) About Eurydice. Available at: http://eacea.ec.europa.eu/education/eurydice/about_eurydice_en.php (accessed 20 December 2014).

Eurydice (2014b) Compulsory education in Europe 2014/2015, facts and figures, November 2014. Available at: http://eacea.ec.europa.eu/education/eurydice/documents/facts_and_figures/compulsory_education_EN.pdf (accessed 20 December 2014).

Eurydice (2014c) The structure of the European education systems 2014/2015, schematic diagrams, November 2014. Available at: http://eacea.ec.europa.eu/education/eurydice/documents/facts_and_figures/EN_2014_15_diagrams_version_finale_pngs.pdf (accessed 20 December 2014).

Eurydice (2015) Eurydice. Better knowledge for better policies. Available at: http://eacea.ec.europa.eu/education/eurydice/contacts_national_units_en.php (accessed 21 December 2015).

Grek S (2009) Governing by numbers: The PISA effect in Europe. *Journal of Education Policy* 24(1): 23–37.

Grek S (2013) Expert moves: International comparative testing and the rise of expertocracy. *Journal of Education Policy* 28(5): 605–709.

Grek S and Lawn M (2009) A short history of Europeanizing education: The new political work of calculating the future. *European Education: Issues and Studies* 41(1): 32–54.

Grek S and Lawn M (2012) *Europeanising Education: Governing a New Policy Space*. Oxford: Symposium Books.

Grek S and Ozga J (2010) Re-inventing public education. The new role of knowledge in education policy making." *Public Policy and Administration* 25(3): 271–88.

Gunter HM, Hall D and Mills C (2014) Consultants, consultancy and consultocracy in education policymaking in England. *Journal of Education Policy* 30(4): 518–539.

Hargadon AB (1998) Firms as knowledge brokers: Lessons in pursuing continuous innovation. *California Management Review* 40(3): 209–227.

Hood CC and Jackson M (1991) *Administrative Argument*. Aldershot: Dartmouth Publishing.

Husén T and Postlethwaite TN (ed.) (1985) *The International Encyclopaedia of Education: Research and Studies*. Oxford: Pergamon.

Husén T and Postlethwaite TN (ed.) (1994) *The International Encyclopedia of Education: Research and Studies*. Oxford: Pergamon.

Jones PW (2004) Taking the Credit: Financing and policy linkages in the education portfolio of the World Bank." In: Steiner-Khamsi G (ed.) *The Global Politics of Educational Borrowing and Lending*. New York: Teachers College Press, pp.188–200.

Kamens DH (2015) Globalisation and the emergence of an audit culture: PISA and the search for 'best practices' and magic bullets. In: Meyer HD and Benavot A (eds) *PISA, Power, and Policy – The Emergence of Global Educational Governance*. Oxford: Symposium Books, pp.117–139.

Knorr Cetina K (1999) *Epistemic Cultures. How the Sciences Make Knowledge*. Cambridge, MA: Harvard University Press.

Lärarnas Tidning (2014) Tunga nej til betyg I årskurs 4 [Heavy no to grades in Year 4]. Available at: http://www.lararnasnyheter.se/lararnas-tidning/2014/11/25/tunga-nej-betyg-arskurs-4 (accessed 1 May 2015).

Latour B and Woolgar S (1979/1986) *Laboratory Life. The Construction of Scientific Facts*. New Jersey: Princeton University Press.

Lindblad S, Pettersson D and Popkewitz TS (2015) International comparisons of school results: A systematic review of research on large scale assessments in education. *Report from the SKOLFORSK project*. Stockholm: The Swedish Research Council. Available at: https://publikationer.vr.se/produkt/international-comparisons-of-school-results-a-systematic-review-of-research-on-large-scale-assessments-in-education/

Lundahl C (2006) *Viljan att veta vad andra vet. Kunskapsbedömning i tidigmodern, modern och senmodern skola* [*To know what others know. Assessment in education in pre-modern, modern, and late-modern times.*] Dissertation, Uppsala University, Sweden.

Lundahl C (2014a) The book of books – encyclopaedic writing in the science of education in the 1980s. In: Nordin A and Sundberg D (eds) *Transnational Policy Flows in European Education*. London: Symposion Books, pp.79–103.

Lundahl C, Hultén M, Klapp A, and Mickwitz L (2015) *Betygens geografi - forskning om betyg och summativa bedömningar i Sverige och internationellt* [*The Geography of grading - Research review on grading and summativ assessments in Sweden and internationally.*] Stockholm: Swedish Research Council.

Lundahl C and Jönsson A (2010) Ännu en gång fuskar regeringen i betygsfrågan [Once again the government cheats in the grading issue]. Available at: http://www.dn.se/debatt/annu-en-gang-fuskar-regeringen-i-betygsfragan/ (accessed 9 July 2016).

Lundahl C and Waldow F (2009) Standardisation and quick languages: The shape-shifting of standardised measurement of pupil achievement in Sweden and Germany." *Comparative Education* 45(3): 365–385.

Lundgren UP (2011) PISA as a political instrument. One history behind the formulating of the PISA programme. In: Pereyra M-A, Kothoff H-G and Cowen R (eds) *PISA Under Examination. Changing Knowledge, Changing Tests, Changing Schools*. Rotterdam: Sense Publisher, pp. 17–30.

Musiał K (2000) *Roots of the Scandinavian Model. Images of Progress in the Era of Modernisation*. Baden-Baden: Nomos Verlagsgesellschaft.

OECD (2013) *Review on Evaluation and Assessment Frameworks for Improving School Outcomes*. Paris: OECD.

OECD (2015) Improving schools in Sweden: An OECD perspective. Available at: http://www.oecd.org/edu/school/improving-schools-in-sweden-an-oecd-perspective.htm (accessed 14 March 2015).

Ozga J, Dahler-Larsen P, Segerholm C, et al. (2011) *Fabricating Quality in Education: Data and Governance in Europe*. London & New York: Routledge.

Peck J and Theodore N (2015) *Fast Policy: Experimental Statecraft at the Thresholds of Neoliberalism*. Minneapolis: University of Minnesota Press.

Petersson O (2013) Svenska politiker har fått sämre beslutsunderlag [Swedish politicians receive weaker evidence for decision making]. Available at: http://www.olofpetersson.se/_arkiv/skrifter/respons2013_sou.pdf (accessed 22 December 2015).

Pettersson D (2008) *Internationell kunskapsbedömning som inslag i nationell styrning av skolan*. Dissertation. Uppsala University, Sweden.

Phillips D (2004) Towards a theory of policy attraction in education. In: Steiner-Khamsi G (ed.) *The Global Politics of Educational Borrowing and Lending*. New York: Teachers College Press, pp.54–67

Prøitz TS (2015) Uploading, downloading and uploading again – concepts for policy integration in education research. *NordSTEP* 2015(1): 70–80.

Ringarp J and Waldow F (2016) From 'silent borrowing' to the international argument-legitimating Swedish educational policy from 1945 to the present day. *Nordic Journal of Studies in Educational Policy* 2(1).

Rizvi F and Lingard B (2010) *Globalizing Education Policy*. London: Routledge.

Robertson SL (2005) Re-imagining and rescripting the future of education: global knowledge economy discourses and the challenge to education systems. *Comparative Education* 41(2): 151–170. DOI: 10.1080/03050060500150922

Rönnberg L, Lindgren J and Segerholm C (2013) In the public eye: Swedish school inspection and local newspapers: Exploring the audit-media relationship. *Journal of Education Policy* 28(2): 178–197.

Schreier M (2012) *Qualitative Content Analysis*. London: SAGE Publications.

Schriewer J (1988) The method of comparison and the need for externalization: Methodological criteria and sociological concepts. In: Schriewer J and Holmes B (eds)*Theories and Methods in Comparative Education*. Frankfurt am Main: Peter Lang, pp. 62–83.

Schriewer J (2003) Comparative education methodology in transition: Towards a science of complexity." In: Schriewer J (ed.) *Discourse Formation in Comparative Education*. Frankfurt am Main: Peter Lang, pp.3–52.

Schriewer J (2014) Neither orthodoxy nor randomness: Differing logics of conducting comparative and international studies in education. *Comparative Education* 50(1): 84–101.

SOU (1977) *Betygen i Skolan* [*The Grades In Schools*]. Stockholm: Statens Offentliga Utredningar.

SOU (2007) *Utredningen om mål och Uppföljning i Grundskolan* (2007). *Tydliga mål och Kunskapskrav i Grundskolan: Förslag till nytt mål- och Uppföljningssystem: Betänkande*. [*Investigation of Clearer Objectives and Better Evaluation in Swedish Curriculum*]. Stockholm: Fritzes.

SOU (2010) *Biologiska Faktorer och Könsskillnader i Skolresultat*. [*Biological Factors and Gender Differences in School Outcomes*]. Stockholm: Statens Offentliga Utredningar.

Steiner-Khamsi G (2004) *The Global Politics of Educational Borrowing and Lending*. New York: Teachers College Press.

Steiner-Khamsi G (2010) The politics and economics of comparison. *Comparative Education Review* 54(3): 323–342.

Steiner-Khamsi G and Waldow F (2012) Policy borrowing and lending. In: Steiner-Khamsi G and Waldow F (eds) *World Yearbook of Education 2012*. New York: Routledge.

Utbildningsdepartementet [Ministry of Education] (2014a) Utbildningsdepartementet. Utredningen om trygghet och studiero i skolan. [Review on conduct and order in schools.] Unpublished.

Utbilningsdepartementet (2014b) En bättre skolstart för alla: bedömning och betyg för progression i lärandet [A better school start for all: Assessment and grading for progression in learning]. Regeringskansliet.

Waldow F (2009) Undeclared imports: Silent borrowing in educational policy-making and research in Sweden. *Comparative Education* 45(4): 477–494.

Zymek B (2003) Domination, legitimacy and education: Max Weber's contribution to comparative education. In: Schriewer J (ed.) *Discourse Formation in Comparative Education*. Frankfurt am Main: Peter Lang, pp. 133–151.

Author biographies

Sverre Tveit (b. 1982) is a university lecturer at the University of Agder, Norway. The paper is part of his PhD study at the University of Oslo. His research interests are primarily in the area of educational assessment, policymaking, policy borrowing and teacher education.

Christian Lundahl (b. 1972) is a professor at Örebro University, Sweden. His research interests are primarily in the areas of educational history, assessment in education and curriculum theory. Lundahl is a member of the research group Education and Democracy.

Appendix 1. A comparison of Eurydice data on educational assessment

In a report to the Swedish Research Council (Lundahl et al., 2015) we generated a comparison of countries' educational assessment and grading systems based on available data from the Eurydice network. The data were generated in fall 2014. Tables 3 and 4 give an overview of the complex information. The tables complement one another. Table 3 overviews the structure of compulsory education including required formal grading. Table 4 overviews the grading scales for the respective countries. Information about the classification of information for each column is given below. Some errors or lost information may occur. Asterisks (*) indicate amendments made to facilitate consistent classification. These are explained below the tables.

Structure of compulsory education systems

Starting age. Students' age when commencing compulsory education. Note that the classification of years/levels in Eurydice may not be fully comparable. Further, different interpretations across the countries in this report may cause variations. Thus, all indication of age for starting school, starting formal marking, differentiation etc. may be somewhat imprecise. However, if the Eurydice information provided is correct, it is unlikely that the descriptions are out by more than a year.

Compulsory years. Number of years of compulsory education (note reservations).

School structure. Lists the type of compulsory structure classified in three types: single structure education systems where there is no parallel education during compulsory years; primary secondary divided education system where there may be merit-based selection when moving from primary to secondary, for example, due to requirements for progressing to next level; and tracked

secondary education systems where secondary students are differentiated according to merit/potential and/or interests.

Differentiation. Details the students' age and year when the education system is differentiated. Differentiation is described with age (year) when differentiated schools/programmes start. Single structure systems and primary secondary divided education system differentiation usually occurs at the conclusion of compulsory education (usually the conclusion of lower secondary education), while tracked secondary education systems occur at the beginning of secondary school.

First certificate. Lists the year and age students get the first official certificate, to the extent this is described. Classified based on year and students' typical age when completing that year. As it is classified based on completion of the year instead of when commencing, student age is up by one year compared with the classifications for 'starting school' and 'grading required'. In some countries, there are matriculation certificates in the forthcoming year; these countries are still classified according to the conclusive year before matriculation.

Grading required. Lists the age and year when teachers' grading is required. This is classified with age when starting the relevant year. This classification should be viewed as indicative, as many countries do not address the issue of formal grading explicitly in their reports.

Grading scales

First grading scale. Lists the first grading scale students meet in school. A significant amount of information is missing. Lists the grading scales that have been identified in the generated data. In many countries (such as the UK countries), students meet grading scales earlier than reported here; however, in many cases Eurydice does not list scales used for classroom assessment. Some countries have different grading scales in primary and secondary education or other ways of combining multiple grading scales. Only the reported grading scales that students meet first are listed here. This should not be taken as an exhaustive reporting of grading scales, as many countries have either not reported such information properly or, as was true in several instances, information was lacking in Eurydice.

Scale levels. Number of grading levels considered, including one failure level (some countries have additional failing levels).

Scale type. Classifies the type of grading scale, to the extent it has been possible to establish, along the following attributes: verbal, verbal and letters, verbal and numerical, numerical, letters and points scale.

Grading behaviour. Addresses whether the countries reported having a legal framework for grading students' behaviour with distinct grades. There are various ways and terms for grading behaviour (order, conduct, diligence). This should not be regarded as an exhaustive classification; it appears that only countries that employ such practices describe them, yet it is not possible to know whether that implies that the others do not have them.

Table 3. Structure of the education systems in Europe.

Country	Starting age	Comp. years	School structure	Differentiation	Certification starts	Grading required
Austria	Age 6	9	Tracked secondary	Age 10 (Year 5)	Year 4 (Age 10)	Age 6 (Year 1)
Belgium (Flemish)	Age 6	9	Tracked secondary	Age 12 (Year 7)	Year 6 (Age 12)	Not described
Belgium (French)	Age 6	9	Tracked secondary	Age 12 (Year 7)	Year 6 (Age 12)	Age 7 (Year 2)
Belgium (German)	Age 6	9	Tracked secondary	Age 12 (Year 7)	Year 6 (Age 12)	Not described
Bulgaria	Age 7	9	Single structure	Age 15 (Year 9)	Year 8 (Age 15)	Not described
Croatia	Age 6*	9	Single structure	Age 14 (Year 9)	Not described	Not described
Cyprus	Age 6*	10	Primary secondary	Age 12 (Year 7)	Year 1 (Age 7)	Age 6 (Year 1)
Czech Republic	Age 6	9	Primary secondary	Age 15 (Year 10)	Year 5 (Age 11)	Not described
Denmark	Age 7*	9	Single structure	Age 16 (Year 10)	Year 9 (Age 16)	Age 14 (Year 8)
Estonia	Age 7*	9	Single structure	Age 16 (Year 10)	Not described	Not described
Finland	Age 7	9	Single structure	Age 16 (Year 10)	Year 9 (Age 16)	Age 14 (Year 8)
France	Age 6	10	Primary secondary	Age 11 (Year 6)*	Year 9 (Age 15)	Age 6 (Year 1)
Germany	Age 6	9	Tracked secondary	Age 10** (Year 5)	Year 9 (Age 15)	Age 7 (Year 2)
Greece	Age 6*	10	Primary secondary	Age 12 (Year 7)	Year 9 (Age 15)	Age 8 (Year 3)
Hungary	Age 6	9	Single structure	Unclear	Year 1 (Age 7)	Age 6 (Year 1)
Iceland	Age 6*	10	Single Structure	Age 16 (Year 11)	Year 10 (Age 16)	Not described
Ireland	Age 6*	10	Tracked secondary	Age 12 (Year 7)	Year 7 (Age 13)	Not described
Italy	Age 6*	10	Inconsistency	Unclear	Year 10 (Age 16)	Age 6 (Year 1)
Latvia	Age 7*	9	Inconsistency	Unclear	Year 9 (Age 16)	Not described
Lichtenstein	Age 6	9	Tracked secondary	Age 11 (Year 6)	Not described	Age 11 (Year 6)
Lithuania	Age 7	9	Primary secondary	Unclear	Year 4 (Age 11)	Age 11 (Year 5)
Luxembourg	Age 6	12	Tracked secondary	Unclear	Year 6 (Age 12)	Age 7 (Year 2)
Malta	Age 5	11	Primary secondary	Unclear	Year 11 (Age 16)	Age 8 (Year 4)
Netherlands	Age 6*	13	Tracked secondary	Age 13 (Year 8)	Age 13 (Year 7)	Not described
Norway	Age 6	10	Single structure	Age 16 (Year 11)	Year 10 (Age 16)	Age 13 (Year 8)
Poland	Age 6*	10	Primary secondary	Unclear	Year 1 (Age 7)	Age 6 (Year 1)
Portugal	Age 6	12	Single structure	Age 15 (Year 10)	Year 6 (Age 12)	Age 9 (Year 4)
Romania	Age 6	11	Primary secondary	Age 15 (Year 10)	Year 9 (Age 15)	Age 6 (Year 1)
Slovakia	Age 6	10	Single structure	Unclear	Year 1 (Age 7)	Age 9 (Year 4)

(Continued)

Table 3. (Continued)

Country	Starting age	Comp. years	School structure	Differentiation	Certification starts	Grading required
Slovenia	Age 6	9	Single structure	Age 15 (Year 10)	Year 1 (Age 7)	Age 8 (Year 3)
Spain	Age 6	10	Single structure	Unclear	Year 10 (Age 16)	Not described
Sweden	Age 7	9	Single structure	Age 16 (Year 10)	Year 9 (Age 16)	Age 12 (Year 6)
Turkey	Age 6*	12	Single structure	Age 15 (Year 9)	Year 4 (Age 11)	Age 7 (Year 1)
UK: England	Age 5	11	Primary secondary	Age 16 (Year 12)	Year 11 (Age 16)	Age 6 (Year 2)
UK: Northern Ireland	Age 4	12	Primary secondary	Age 16 (Year 12)	Year 11 (Age 16)	Age 5 (Year 2)
UK: Scotland	Age 5	11	Primary secondary	Age 16 (Year 12)	Year 11 (Age 16)	Not described
UK: Wales	Age 5	11	Primary secondary	Age 16 (Year 12)	Year 11 (Age 16)	Age 6 (Year 2)

Comp.: compulsory

Table 4. Grading scales in Europe based on information from Eurydice.

Country	First grading scale	Scale levels	Scale type	Grading behaviour
Austria	Very good (1), Good (2), Satisfactory (3), Sufficient (4), Insufficient (5)	5	Verbal and numerical	Not described
Belgium (Flemish)	Not described	Not described	Not described	Not described
Belgium (French)	Not described	Not described	Not described	Not described
Belgium (German)	Not described	Not described	Not described	Not described
Bulgaria	Excellent (6), Very good (5), Good (4), Fair (3) Poor (2)	6*	Verbal and numerical	Not described
Croatia	Excellent, Very good, Good, Sufficient, Insufficient.	5	Verbal	Not described
Cyprus	Lower secondary: A=Excellent, B=Very Good, C=Good, D=Almost Good, E=Fail	5	Verbal and letters	Not described
Czech Republic	Not described	Not described	Not described	Not described
Denmark	12, 10, 7, 4, 02, 00, –3	6**	Numerical	Not described
Estonia	Very good (5), Good (4), Satisfactory (3), Poor (2), Weak (1)	5***	Verbal and numerical	Not described
Finland	Excellent (10), Very Good (9), Good (8), Satisfactory (7), Moderate (6), Adequate (5), Fail (4)	7	Verbal and numerical	Not described
France	20-1.	20	Not described	Not described
Germany	Very good (1), Good (2), Satisfactory (3), Adequate (4), Weak (5), Very Weak (6)	6	Verbal and Numerical	50% of the states
Greece	Years 3 and 4: Excellent (A), Very good (B), Good (C), Fairly Good (D)	4****	Verbal and letters	Not described
Hungary	Very good (5), Good (4), Satisfactory (3), Pass (2), Fail (1)	5	Verbal and Numerical	Conduct, diligence
Iceland	Honours (9.00–10.00), First Class (7.25–8.99), Second Class (6.00–7.24), Third Class (5.00–5.99), Fail (0.00–4.99)	5	Verbal and Numerical	Not described
Ireland	Not described	Not described	Not described	Not described
Italy	10-1 (6 is pass)	Not described	Not described	Behaviour
Latvia	10 (Excellent), 9 (Excellent), 8 (Very good), 7 (Good), 6 (Almost Good), 5 (Satisfactory), 4 (Almost Satisfactory), 3 (Weak), 2 (Very Weak), 1 (Very, Very Weak)	7*****	Verbal and numerical	Not described
Lichtenstein	Not described	Not described	Not described	Not described
Lithuania	Not described	Not specified	Not specified	Not described
Luxembourg	Upper secondary: Excellent (> 52 points), Assez bien (36–39 points), Bien (40–47 points), Très bien (48–51 points)	Not described	Points scale	Not described
Malta	1-100	Not described	Points scale	Not described
Netherlands	Not described	Not described	Not described	Not described
Norway	6 (Exceptional), 5 (Very High), 4 (High), 3 (Fair), 2 (Low), 1 (Very Low)	6	Verbal and numerical	Order and conduct
Poland	Excellent (6), Very good (5), Good (4), Satisfactory (3), Acceptable (2), Poor (1)	6	Verbal and numerical	Behaviour (conduct)

(Continued)

Table 4. (Continued)

Country	First grading scale	Scale levels	Scale type	Grading behaviour
Portugal	5, 4, 3, 2, 1 (3 minimum to pass)	4	Not specified	Not described
Romania	Primary: Very good, Good, Sufficient, Insufficient	4	Verbal	Behaviour
Slovakia	Excellent (1), Laudable (2), Good (3), Satisfactory (2), Fail (5)	5	Verbal and numerical	Behaviour
Slovenia	5, 4, 3, 2, 1	5	Numerical	Not described
Spain	Not described	Not described	Not described	Not described
Sweden	A, B, C, D, E, F	6	Letters	Not described
Turkey	Not described	9	Not described	Not described
UK: England	General Certificate of Secondary Education: 9-8-7-6-5-4-3-2-1	9	Numerical	Not described
UK: Northern Ireland	General Certificate of Secondary Education: 9-8-7-6-5-4-3-2-1	9	Numerical	Not described
UK: Scotland	Not described	9	Not described	Not described
UK: Wales	General Certificate of Secondary Education: 9-8-7-6-5-4-3-2-1	9	Numerical	Not described

Reservations, Tables 3 and 4:
Starting age.
*Croatia: there is a compulsory pre-school year; however, it is not counted as the first school year.
*Cyprus: school starts at age 5 years + 8 months. There is a compulsory pre-school year; however, it is not counted as the first school year.
*Denmark: there is one compulsory pre-school year; however, it is not counted as the first school year. The optional 10th year before upper secondary is not counted, thus the first secondary year is listed as Year 10.
*Estonia: there is a compulsory pre-school year; however, it is not counted as the first school year.
*Greece: there is a compulsory pre-school year; however, it is not counted as the first school year.
*Iceland: there is a compulsory pre-school year; however, it is not counted as the first school year.
*Ireland: there is a compulsory pre-school year; however, it is not counted as the first school year.
*Italy: there is a compulsory pre-school year; however, it is not counted as the first school year.
*Latvia: there are two compulsory pre-school years; however, they are not counted as first school years. The optional 10th year before upper secondary is not counted, thus the first secondary year is listed as Year 10.
*Netherlands: there is a compulsory pre-school year; however, it is not counted as the first school year.
*Poland: there is a compulsory pre-school year; however, it is not counted as the first school year.
*Turkey: school starts age 6 years +6 months. There is a compulsory pre-school year; however, it is not counted as the first school year.
Compulsory years.
*Germany: in 12 states, there are nine compulsory years; in five states, there are 10 compulsory years.
Differentiation.
*France: there are many school types, and thus it is highly complex to describe how students are differentiated based on merit.
**Germany: in Berlin and Brandenburg, secondary levels start at age 12 (Year 7).
Grading levels.
*Bulgaria: failing level is not described.
**Denmark: one additional failure level.
***Estonia: failure level is not addressed.
****Greece: Years 1–6 have four levels; Years 7–9 have five levels.
*****Latvia: additional three failing levels.

IV

Sverre Tveit[1]

(Trans)national Trends and Cultures of Educational Assessment: Reception and Resistance of National Testing in Sweden and Norway during the Twentieth Century

The twentieth century was a time of increasing international relations in education policy and research. Lawn (2013) observes that "the field of education was riven with the problems of the expansion of secondary education, selection processes and school outcomes."[2] Research centres and international projects became central nodes for solving policy problems in national education systems. Educational measurement thus became "a defining element of the governing of education."[3] The growth of intelligence expertise in the United States during the interwar years paved the way for new approaches to educational measurement for multiple purposes that were circulated through new institutions, such as United Nations Educational, Scientific and Cultural Organisation (UNESCO), the International Association for the Evaluation of Educational Achievement (IEA), and the Organisation for Economic Co-operation and Development (OECD).

The objective of this paper is to outline analytical-conceptual frameworks for understanding transnational trends with respect to the various roles of educational assessment emphasised in national assessment instruments, which emerged in concert with the increased international collaboration during the twentieth century described above. The first framework identifies three trends that can be related to transnational research and policy endeavours: First, the *meritocracy* trend, focusing on fair certification and selection procedures for individual

---

1 This chapter reports in-part on archive investigations into Columbia University's Rare Book & Manuscript Libriary, which were undertaken with support from the Ryoichi Sasakawa Yong Leaders Fellowship Fund (Sylff) scholarship provided by the University of Oslo and the Sylff Research Abroad scholarship provided by the Tokyo Foundation. The author is also indebted to Dr. Thomas Hatch at Teachers College, Columbia University, who facilitated the research visit.

2 Martin Lawn, "Voyages of Measurement in Education in the Twentieth Century: Experts, Tools and Centres," *European Educational Research Journal* 12, no. 1 (2013): 109, doi:10.2304/eerj.2013.12.1.108.

3 Ibid.

students, was emphasised in international research projects such as the International Examinations Inquiry (IEI) in the 1930s. Second, the *accountability* trend, which places more emphasis on the governing of education systems and their role in global competition among national states, became more prominent when comparative testing programmes were organised in fixed cycles from the 1990s onwards. Third, the *Assessment for learning* trend, emphasising the role of assessment instruments and procedures in supporting student learning, emerged at the change of the millennium. While researchers and policymakers may emphasise this third trend as a reaction (or in opposition) to the effects associated with the meritocracy and accountability trends, the OECD also embraced it as a key strategy of the accountability policies it advocates. As such, *Assessment for learning* may by some be perceived as a trend that reacts and is in opposition to the accountability trend, while for others it is subordinated to the accountability trend's emphasis on strategies for improving countries' educational outcomes. In sum, these three transnational trends of educational assessment have shaped the roles of educational assessment emphasised in countries' national assessment instruments worldwide.

The second framework utilises Hopmann's (2003) distinction between *process*- and *product*-controlled education systems, and relates it to modes of determining students' level of attainment.[4] It is developed to illustrate how product-controlled education systems were more receptive to the accountability trend's quest for measurable outcomes as the basis for governing education because its meritocratic instruments had already been adapted to new psychometric principles. Process-controlled education systems, on the other hand, resisted psychometric approaches to measure outcomes until the PISA shock paved the way for such tests as the basis for governing education in many countries. These two different cultures of certifying and governing learning and instruction are labelled the (American product-controlled) testing tradition and the (continental European process-controlled) examination tradition respectively.

By analysing Sweden and Norway's participation in large scale international assessments, and investigating second hand literature and archive documents that capture these developments, the paper demonstrates how the three transnational trends of educational assessment shaped the emphasis on roles of educational assessment in the countries' national assessment instruments, and how this can be related to the countries' different testing and examinations cultures. With the

4 Stefan T. Hopmann, "On the Evaluation of Curriculum Reforms," *Journal of Curriculum Studies* 35, no. 4 (2003), doi: 10.1080/00220270305520.

examples of Sweden and Norway, the chapter illuminates how different engagement with transnational research projects and trends of educational assessment, and different testing and examination cultures, shaped the accumulation of purposes associated with contemporary national assessment instruments. Conclusively the chapter discusses how the differences between these two Scandinavian countries may be illustrative of wider patterns in European countries' cultures — and their reception and resistance towards transnational trends — of educational assessment.

Analytical Framework: Transnational Trends of Educational Assessment

By integrating theoretical conceptualisations of the purposes of educational assessment in the research literature with an empirical investigation of policy documents in Norway and Sweden, I identify three principal *roles of educational* assessment, that can be associated with the process of determining students' level of attainment in national education systems: Educational assessment used to *certify*, to *govern* and to *support* learning and instruction (Table 1).[i]

However, the roles of assessment instruments cannot be understood through contemporary analyses alone. Contemporary use should be understood in view of how the assessment instruments emerged. While these developments are largely due to domestic factors, they are also a product of transnational influence. Nordin and Sundberg discuss how UNESCO, the World Bank, the OECD and the European Union "have come to play an increasingly important role in the construction of transnational policy arenas, as resourceful actors working together, forming powerful discourse coalitions that influence and to some extent even govern national reforms."[5] Issues that have traditionally been perceived as national in character, such as educational assessment, has also become relevant to the transnational sphere.

To come to terms with what has sparked or influenced changes in the use of national assessment instruments, and with the accumulation of purposes in contemporary policies, I outline the three previously mentioned transnational trends of educational assessment. As will be showed, the *meritocracy*, *accountability*, and *assessment for learning* trends are related to transnational research projects (IEI), research agencies (IEA) and policy agencies (OECD). Each of these trends are elaborated below.

---

5 *Nordin, Andreas,* and *Daniel Sundberg*, "The Making and Governing of Knowledge in the Education Policy Field," in *Transnational Policy Flows in European Education*, eds. Andreas Nordin, and Daniel Sundberg, (Oxford, UK: Symposium Books, 2014), 14.

Meritocracy, a term initially coined by Michael Young (1958) in the satiric text *The Rise of Meritocracy,* characterises the change from a society in which social status is ascribed by birth (aristocracy) to one in which social status depends on individuals' achievements.[6] Young portrayed the twentieth century as obsessed by a (utopic) vision of developing assessment procedures that would distinguish between candidates based on their achievement rather than their social status. Thus, while meritocracy has come to mean the notion of (objective) merit-based qualification, Young's point was that this was a utopia because the elite control the procedures. Nevertheless, the term meritocracy is now commonly associated with a desired principle for fair distribution of educational opportunities in democratic education systems.[7]

Examinations and tests are key institutions of meritocratic education systems, set up to facilitate fair competition. Concerns over the instruments' effectiveness in this regard increased in the first decades of the twentieth century. The International Examinations Inquiry (IEI) of the 1930s was a notable international project that investigated the validity and comparability problems of the tests used for certification and selection purposes. When established, the IEI study included the United States, Scotland, England, France, Germany and Switzerland. It later grew to include Norway, Sweden, and Finland as well. The researchers dealt with the expansion of secondary education and the determination of "the most effective way of examining pupils for entry into the secondary school."[8] The project served as an arena for exchanging experiences with research on the participating countries' assessment instruments. As elaborated further below, it became a node for the exchange of new psychometric approaches to educational assessment.

Accountability became a key focus of international research projects in the second half of the twentieth century, which saw a large increase in

6 Michael Young, *The Rise of Meritocracy* (London, UK: Penguin Books, 1958). The etymological origin is the Latin word 'meritum', which means 'due reward' and is related to the verb 'mereri', 'to earn, deserve'.

7 Gro H. Aas, "Likhet uten solidaritet? Idéhistoriske studier av karakterer I utdanning og meritokrati" (PhD diss., University of Gothenburg, 2006).

8 Martin Lawn, ed., *An Atlantic Crossing? The Work of the International Examination Inquiry, its Researchers, Methods and Influence* (Oxford, UK: Symposium Books, 2008), 7.

performativity-oriented policies[9] and "governing by numbers."[10] Initiatives from global and European actors led to national discussions about what is required of a nation and its inhabitants for excelling in international competition. When national states' human capital takes precedence over products and services as the key factor in economic success, it legitimises external involvement in national education systems.[11] Supranational agencies[12] have therefore made politicians accountable not only to their respective populations but also to European and global standards.[13]

While this development in the contemporary discourse is largely associated with the OECD and the PISA tests, the emphasis on comparing educational outcomes started with studies initiated by UNESCO and the IEA. While the IEA became a legal entity in 1967, the scholarly collaboration dates back to 1958 when a group of scholars, educational psychologists, sociologists, and psychometricians met at the UNESCO Institute of Education. The two first studies undertaken in the 1960s (the Pilot Twelve-Country Study, 1960; the First International Mathematics Study, 1964) included 12 countries. While the emphasis was on a range of subjects in the early 1970s (the Six-Subject Study, 1970–71), the emphasis became more concentrated on mathematics, science, and reading from the 1980s onwards.

The basic idea of the IEA's founders was that different national practices could "lend themselves to comparisons that would yield new insights into the determinants of educational outcomes, servicing as a basis for the improvement of the quality of education."[14] The 1990s saw an increase in global influence on Eastern European and developing countries as well. Following the declaration of the World Conference on Education for All,[15] several less-developed countries

9 Stephen J. Ball, "The Teacher's Soul and the Terrors of Performativity," *Journal of Education Policy* 18, no. 2 (2003), doi: 10.1080/0268093022000043065.

10 Sotiria Grek, et al., "National Policy Brokering and the Construction of the European Education Space in England, Sweden, Finland and Scotland," *Comparative Education* 45, no. 1 (2009), doi: 10.1080/03050060802661378.

11 Eva Forsberg, "Utbildningens Bedömningskulturer I Granskningens Tidevarv," *Utbildning & Demokrati* 23, no. 3 (2014): 53–76.

12 Roger Dale, "Globalisation, Knowledge Economy and Comparative Education," Comparative Education 41, no. 2 (2005), doi: 10.1080/03050060500150906.

13 Grek et al., "National Policy Brokering."; Tine S. Prøitz, "Uploading, Downloading and Uploading Again — Concepts for Policy Integration in Education Research," *Nordic Journal of Studies in Educational Policy* 1 (2015), doi:10.3402/nstep.v1.27015.

14 Lawn, "Voyages of Measurement," 108.

15 UNESCO, "World Declaration on Education for All: Meeting basic needs" (adopted by the World Conference on Education for All, New York, NY: UNESCO, 1990).

embarked on national testing programmes, using expertise developed from cross-Atlantic research collaborations.[16] Global agencies' increased emphasis on national testing as governing instruments coincided with the IEA launching new testing programmes in mathematics, science, and reading, now known as TIMSS (Trends in International Mathematics and Science Study), in 1995 and PIRLS (Progress in International Reading Literacy Study) in 2001. Both were to be undertaken cyclically (every fourth and fifth year, respectively), with more emphasis on facilitating comparisons measures between countries and over time. In the 1990s, the OECD also began its work on its Programme for International Student Assessment (PISA), with the first tests undertaken in 2000, to follow every third year thereafter. The PISA studies radically changed the premises of policy legitimation and education governance globally,[17] causing a "manic search for best practices."[18] In summary, while the IEA's founders were already concerned with tests' governing role in the 1960s, the cyclic use of IEA tests from the 1990s, followed by the OECD's PISA tests in the new millenium, defined the breakthrough of the accountability trend of educational assessment.

Assessment for learning emerged as a new policy area in tandem with (and partly in opposition to) governments and international agencies' increased focus on accountability measures. Sparked by meta-studies that reported impressive effect sizes and compelling arguments for the effectiveness of *formative assessment,*[19] several countries implemented new policies called *Assessment for learning* or similar.[20] In response to the 'standards crisis', governments saw the potential of

16 Thomas Kellaghan, "The Globalisation of Assessment in the 20th Century," *Assessment in Education: Principles* 8, no. 1 (2001), doi: 10.1080/09695940120033270.

17 Heinz-Dieter Meyer and Aaron Benavot, "Introduction," in *PISA, Power, and Policy: The Emergence of Global Educational Governance*, ed. Heinz-Dieter Meyer and Aaron Benavot (Oxford, UK: Symposium Books, 2015).

18 David H. Kamens, "Globalisation and the Emergence of an Audit Culture: PISA and the Search for 'Best Practices' and Magic Bullets," in *PISA, Power, and Policy: The Emergence of Global Educational Governance*, ed. Heinz-Dieter Meyer and Aaron Benavot (Oxford, UK: Symposium Books, 2015), 137.

19 Paul Black and Dylan Wiliam, "Assessment and Classroom Learning," *Assessment in Education* 5, no. 1 (1998), doi:10.1080/0969595980050102; John Hattie, *Visible Learning* (London, UK: Routledge, 2008).

20 For Norway, see Therese N. Hopfenbeck, Maria Teresa Flórez Petour and Astrid Tolo, "Balancing Tensions in Educational Policy Reforms: Large-Scale Implementation of Assessment for Learning in Norway," *Assessment in* Education: Principles 22, no. 1 (2015); for Sweden, see Anders Jonsson, Christian Lundahl and Anders Holmgren, "Evaluating a Large-Scale Implementation of Assessment for Learning in Sweden,"

formative assessment to raise standards through slogans such as "formative use of summative tests."[21] The much-quoted Black and Wiliam review article "Assessment and Classroom Learning"[22] can be perceived as a milestone in the emergence of a greater emphasis on the formative use of tests and teachers role in assessment at the turn of the millennium. In 2002, the OECD's "What Works in Innovation in Education programme" gave emphasis to studies that reported formative assessment to produce educational gains "among the largest ever reported for educational interventions."[23] The book "Formative Assessment: Improving Learning in Secondary Classrooms"[24] featured exemplary cases from secondary schools in eight countries and reviewed research publications in German and French. Assessment for learning and formative assessment policies were key components of OECD's "Review on Evaluation and Assessment Frameworks for Improving School Outcomes" which included 14 countries.[25] The OECD has also taken the role of reviewing countries' "*Assessment for learning*" programmes.[26]

As shown in Table 1, these developments can be viewed as three transnational trends that have influenced countries' use of educational assessment instruments. The years listed do not indicate an exclusive emphasis on the respective purpose but, rather, are the time when countries' policies and associated instruments *accumulated* these educational assessment purposes. The three transnational trends can be linked to three principal roles of educational assessment, and associated

---

Assessment in Education: Principles, Policy & Practice 22, no. 1 (2015), doi:10.1080/0969594X.2014.970612.

21 Wynne Harlen, "On the Relationship Between Assessment for Formative and Summative Purposes," in *Assessment and Learning*, ed. John Gardner (London, UK: Sage, 2006), doi: 10.4135/9781446250808.n6.

22 Black and Wiliam, "Assessment and Classroom Learning."

23 "Centre for Educational Research and Innovation (CERI) — What Works," OECD, accessed July 21, 2015. http://www.oecd.org/edu/ceri/centreforeducationalresearchandinnovationceri-whatworks.htm.

24 OECD, *Formative Assessment: Improving Learning in Secondary Classrooms* (OECD, 2005).

25 "OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes — Country Reviews," OECD, accessed July 22, 2015. http://www.oecd.org/edu/school/oecdreviewonevaluationandassessmentframeworksforimprovingschooloutcomescountryreviews.htm.

26 See, e.g., Therese N. Hopfenbeck, Astrid Tolo, Maria Teresa Florez and Yasmin El Masri, "Balancing Trust and Accountability? The Assessment for Learning Programme in Norway," 2013, accessed May 21, 2015. http://www.oecd.org/edu/ceri/Norwegian%20GCES%20case%20study%20OECD.pdf.

processes of determining students' level of attainment, that has been identified in a review of research literature and an empirical investigation of contemporary policy documents in Sweden and Norway.[27] As described in Table 1, these are called: to *certify*, to *govern* and to *support* learning and instruction.[ii]

Table 1: Roles and Transnational Trends of Educational Assessment

Process	To determine educational goal (or standard) attainment		
Role	To *certify* learning and instruction	To *govern* learning and instruction	To *support* learning and instruction
Level	Student and teacher level (teachers' grading, exit examinations)	Organisational level (schools, municipalities, national states)	Student and teacher level (classroom assessment)
Institutional practice	To identify and report the final level of attainment (a grade/mark, examination); used for certification and selection for further education and professional life	To evaluate (aggregated) student attainment data; used to (a) inform decision makers' quality development efforts; and (b) to control application of curricula and regulations	To identify and communicate gaps between the current and desired attainment levels; used to inform learning and instruction strategies
Transnational Trends	Meritocracy (1930s→)	Accountability (1990→)	Assessment for learning (2000→)
Transnational research projects	International Examinations Inquiry (IEI), 1933–1938	**IEA TIMSS**: 1995, 1999, 2003, 2007, 2011, 2015, 2019 **IEA PIRLS**: 2001, 2006, 2011, 2016 **OECD PISA**: 2000, 2003, 2006, 2009, 2012, 2015, 2018	OECD, 2005 OECD, 2013

In the next section, I sketch the emergence of examinations and tests as certification and governing instruments in Europe and the United States in the nineteenth and

27 Sverre Tveit, "Ambitious and Ambiguous: Sverre Tveit, "Ambitious and Ambiguous: Shifting Purposes of National Testing in the Legitimation of Assessment Policies in Norway and Sweden (2000–2017)," *Assessment in Education: Principles, Policy & Practice* (forthcoming, 2018)

twentieth century. Furthermore, I explain how the roles of Sweden's and Norway's contemporary national assessment instruments reflect the different reception of (American) psychometric approaches to educational assessment, which was mediated both through the meritocracy and accountability trends.

The Emergence of Examinations and Tests in Europe and the United States in the Nineteenth and Twentieth Centuries: Process and Product Control

In the seminal volume *The Measure of Merit*, John Carson[28] describes how the French and American republics responded in different ways to the problem of balancing equality and difference as their education systems expanded from the 1750 to 1940. Combined with Stefan Hopmann's distinction between process- and product-controlled education systems,[29] these perspectives offer a framework for coming to terms with how national assessment instruments (i.e. examinations and tests) emerged as certification and governing instruments in continental Europe[30] and the United States in the nineteenth and twentieth centuries. Hopmann defines *process* and *product control* as two fundamentally different ways of steering the education system through educational assessment,[31] which I contend in part can explain the different emphases on *professional (social) judgement* versus *external (objective) measurement* procedures to facilitate the validity and comparability of assessments. Table 1 envisions the relationship between process- and product-control and the assessment instruments used to govern the education system and its certification procedures.

---

28 John Carson, *The Measure of Merit: Talents Intelligence, and Inequality in the French and American Republics, 1750–1940* (Princeton, NJ: Princeton University Press, 2007).

29 Hopmann, "On the Evaluation of Curriculum Reforms".

30 The qualification 'continental' is used as the premises of process- and product-control in the United Kingdom is more comparable to that of the United States than to e.g. Germany, France and the Scandinavian countries. In contemporary policies, the United Kingdom can be perceived as a blend of the examination and test cultures, by using standardized yet more essay-based tests and the use of external markers, which is different from the largely multiple-choice dominated testing in the United States. This notion of a blended examination and test tradition that mixes the emphasis on external (objective) measurement with professional judgments is not elaborated further in this chapter.

31 Ibid.

Table 2: Relationship between Process- and Product-Control and the Emphasis on professional judgments vs. external measurement[32]

Curriculum steering	**Process-control**	**Product-control**
Premises for controlling the curriculum and teachers:	The national curriculum provides guidelines to teachers, who are recognised as qualified through national teacher education.	The school sector is divided between private and public providers, with no unified concept of teacher education. Thus, the emphasis is on external product control instead.
	↕	↕
Assessment instruments	**EXAMINATIONS**	**TESTS**
Assessment instruments used to govern the education system and its certification procedures rely on:	**Professional (subjective) judgement:** Members of the profession control each other's assessments to facilitate the validity and comparability of assessments.	**External (objective) measurement:** Standardized tests developed by measurement experts facilitate the validity and comparability of assessments.

The (European) Examination Culture

Hopmann (2003) observes that many European countries introduced new ways of controlling and evaluating schools in the nineteenth and twentieth century: teachers were licenced to teach according to their own standards but within centralised guidelines.[33] The teaching profession gained more influence over the centralised guidelines and the definitions of what was considered adequate student attainment, which was reflected in the profession's control over examination procedures. Such *process-control*, Hopmann argues, characterised most of continental Europe. Jarning and Aas note that the *Examen Artium* in Norway and Denmark (legislated in 1809) and the *Studenteksamen* in Sweden and Finland (legislated in 1824) are the functional equivalents of the German *Abitur* and the French

32 The distinction between process and product control was developed by Hopmann (2003), while the distinction between concepts of merit was developed drawing on Carson (2007).

33 Hopmann, "On the Evaluation of Curriculum Reforms."

Baccalauréat.[34] They belong to a pattern of key national educational institutions of liberal modernity.[35]

According to Hopmann, the role of examinations in process-controlled education systems has its roots in post-Napoleonic Prussia.[36] From the 1820s onwards, this system of curriculum control "diffused through most of continental Europe."[37] This system is based on the principle of the state providing general curriculum guidelines that outline what to teach, combined with prescriptions for who is qualified to teach (having passed required teaching examinations), but leaves the pedagogical or methodological freedom to the local teaching staff or school.[38] Hopmann continues: "This open system of process control enhanced the independence of the teaching profession, which then turned against all other forms of external school evaluation and control, denouncing them as not being professionally grounded."[39] "Passing the final internal exams of one type of school became enough to gain access to the following stages."[40] During the expansion of the education system in the twentieth century, examinations were a tool for controlling an otherwise largely autonomous teacher profession.[41]

Carson observes that, unlike the American republic (discussed below), the French adopted a national, universal, and comprehensive approach to education with rigorous examinations, relying on expert judgments to determine which students should move up in the system.[42] French psychologists invented the modern intelligence test that the Americans later embraced — the Binet-Simon intelligence scale. However, French administrators were ambivalent about employing the new technology in their meritocratic procedures and preferred to assess individuals on the basis of methods that relied on expert judgement: "Rigorous examinations determined who could move up, with the goal of ensuring that the most talented

34 Harald Jarning and Gro H. Aas, "Between Common Schooling and the Academe: The International Examinations Inquiry in Norway, 1935–1961," in *An Atlantic Crossing? The Work of the International Examination Inquiry, its Researchers, Methods and Influence*, ed. Martin Lawn (Oxford, UK: Symposium Books, 2008).

35 Detlef Müller, Fritz K. Ringer and Brian Simon, *The Rise of the Modern Educational System* (Cambridge, UK: Cambridge University Press, 1986).

36 Hopmann, "On the Evaluation of Curriculum Reforms."

37 Ibid., 470.

38 Ibid., 469.

39 Ibid., 470.

40 Ibid.

41 Christian Lundahl and Sverre Tveit, "Att legitimera nationella prov i Sverige och i Norge – en fråga om profession och tradition," *Pedagogisk forskning i Sverige,* no. 4–5 (2014).

42 Carson, *The Measure of Merit*.

received the best education and became the core of the nation's technocratic elite."[43] Although substantial amendments were made during the expansion and modernisation of education throughout the twentieth century, the principle remains in many European countries: the responsibility for certifying education is undertaken by teachers, under modest state control through examination systems.

The (American) Testing Culture

In countries where the public-school sector or the teaching profession "failed to secure the same prominence as it did in much of continental Europe"[44], a tradition of product control emerged instead. The history of schooling in the former British Empire offers many examples of traditions of product control. United States is the most prominent example, as education was a local affair and no national system of teacher education existed.[45] In line with Hopmann's observations, Carson[46] captures how different traditions for determining merit emerged in France and the United States as the education systems expanded over the course of the nineteenth and twentieth centuries. The American republic however, put more weight on personal attributes than on formal education and embraced intelligence tests as a means of social advancement or distinction. By the 1920s and 1930s, distinctly different ways of understanding differences in mental abilities had emerged. The technology of intelligence testing that Binet and Simon initiated was employed by the Americans for military recruitment during World War I. In the interwar period, testing underwent an enormous boom. What we now know of as the American SAT tests were first used for college admission in 1926. The methodological approaches that started with intelligence testing in the early twentieth century emerged to become comprehensive methods for educational measurement over the next decades.

According to Brookhart "almost all summative assessment and grading in schools were based on teacher judgement" in the United States until the minimum competence movement of the 1970s and 1980s.[47] Brookhart contends that studies of teacher judgments undertaken in the early twentieth century nevertheless had,

43 Ibid., 4.

44 Hopmann, "On the Evaluation of Curriculum Reforms," 471.

45 Ibid.

46 Carson, The Measure of Merit.

47 Susan M. Brookhart, "The Use of Teacher Judgement for Summative Assessment in the USA," *Assessment in Education: Principles, Policy & Practice* 20, no. 1 (2013): 70, doi:10.1080/0969594X.2012.70317.

> "Set the stage for a distrust of teacher judgement of the quality of students' work, a perspective that has been typical of the attitudes towards teacher judgement in the United States ever since [….]. The 'new science' of education swept in with the solution to the problem of unreliable teacher judgement: standardized, objective testing of student achievement."[48]

These increasingly found their way into classrooms as the public trust in the quality of education fell post World War II. Around this time, viewed from the United States, the developments in the Soviet Union represented the anti-thesis to democratic education. A House Committee on "un-American activities'" led by Richard Nixon argued that the Soviet system was set up to give loyal teachers "new and extreme authority over their pupils, who in turn have become cowed, uniformed puppets."[49] A distinction was drawn between 'training' and 'education', where the latter was held to foster independent thinkers as opposed to 'trained puppets.' Tröhler noted that the Soviet Union's 1957 launch of the Sputnik satellite — the first human-made object to orbit the earth — was a shock for the Americans, who had predicted the failure of the education system of the Soviet Union.[50] It "triggered an educational offensive designed to serve both the military and the economic development."[51] In 1958, President Eisenhower introduced the first national law in education, the National Defense Education Act. This was the start of a shift from education viewed as a cultural system to a view of it as a production system.

The American constitution does not allow the federal government to mandate curriculum and teaching reforms. Instead, the states and local school districts were motivated to undertake reform through funding incentives, first introduced by President Lyndon B. Johnson in the 1965 Elementary and Secondary Education Act (ESEA). The federal government could not govern directly, yet it "at least wanted to see what effects its incentives had, and for this purpose a test instrument had to be developed."[52] While the administration of education was and remains a local affair in the United States, this federal involvement marked a shift from input to output steering at the local level that further enhanced the emphasis

48 Ibid., 72.

49 House Committee on Un-American Activities, Title, 57, quoted in Daniel Tröhler, "Truffle Pigs, Research Questions, and Histories of Education," in *Rethinking the History of Education: Transnational Perspectives on Its Questions, Methods, and Knowledge*, ed. Thomas S. Popkewitz (New York, NY: Palgrave Macmillan US, 2013), 114.

50 Ibid., 145.

51 Ibid., 145.

52 Ibid., 150.

on measuring education outcomes of both individual students and schools.[53] Thus — in comparison with continental Europe — the American emphasis on product-control was further propelled by output steering related to funding provided via the federal budget through agreements in the national education acts. ESEA has been revised several times since. In the past decade it has been known as No Child Left Behind, with increased emphasis on holding schools and teachers accountable for student outcomes.

Parallel to the developments towards output steering that was enhanced with the federal involvement in education in the United States, American scholars such as Ralph Tyler and Benjamin Bloom were central in the development of new approaches to curriculum programs and instruction methods. Tyler's landmark eight-year study investigating the effects of progressive education methods in high schools in the 1930s produced a set of principles for educational program evaluation. Bloom's theories on behavioural objectives, master learning and measurement-driven instruction "pushed Tyler's principles of evaluating broad learning outcomes at the school and programme level to the level of fine-grained, classroom lesson objectives."[54] Madhaus noted that Ralph Tyler's contributions to testing and to curriculum development and its evaluation "were both a product and a victim of the times," as it coincided with the rise of behaviourism.[55] Theories for 'programmed instruction' emerged drawing on theories of the 'teaching machine' proposed by behavioural psychologist B.F. Skinner in the mid-1950s. "The general idea behind programmed instruction and teaching machines was that knowledge can be split in many easy-to-learn, small, and consecutive steps to be learned individually."[56] Tyler's and Bloom's writings were on hand when large-scale program evaluation was mandated in order to qualify for federal funding. "The adaptation of both Tyler's and Bloom's works to the needs of the time changed the way people understood them over the ensuing four decades."[57]

In the 1970s the minimum competency movement sparked the use of standardized tests. This was a reaction to the dissatisfaction with public education. By

53 National Assessment of Educational Progress (NAEP) began with a grant from the Carnegie Corporation in 1964. Administrated by a centre of the US Department of Education, the first NAEP tests were administrated in 1969.

54 Brookhart, "The Use of Teacher Judgement," 70.

55 George F. Madhaus, "Ralph Tyler's Contribution to Program Evaluation," in *Evaluation Roots. A Wider Perspective of Theorists' Views and Influences*, 2nd edition, ed. Marvin C. Alkin (Los Angeles, CA: Sage, 2013), 162.

56 Tröhler, *Truffle Pigs*, 148.

57 Madhaus, "Ralph Tyler's Contribution," 162.

1980 minimum competency testing for reading and mathematics was required in 29 states. These were external to the classroom and neither made nor scored by teachers. When it became clear that minimum competency testing lowered expectations to meet the minimum requirements it was succeeded by the educational reform movement of the 1980s and 1990s, later known as the 'standards movement'. The national commission report, "A Nation at Risk" (National Commission on Excellence in Education, 1983), advocated "rigorous and measurable standards and high expectations, a commitment to both excellence and equity, and recommended state and local use of standardized achievement tests."[58] This approach was further expanded with the No Child Left Behind Act of 2002, which mandated annual standards-based tests in grades 3–8 and once during high school.

In summary, the developments in the United States started without a national education system and associated teacher profession, which paved the way for a product-controlled education system. In lack of national and state structures for the organisation of schooling and teacher education from the outset, the premises of process control were not present in the United States. As studies showed poor inter-rater reliability of teacher judgements, psychometric tests gained preference as certification instruments as the basis for college admission. Public dissatisfaction with the standard of education from the 1960s onwards, further propelled the product-controlled education system as federal investment in public education was tied with psychometric measures of student outcomes in order to hold schools and teachers accountable.

While there are many reasons for the increased emphasis on standardized testing in the United States in the twentieth century,[59] the above brief outline of the (European) examination and (American) testing cultures establish that they, in part, can be explained by different premises of process- and product-control. The next section discusses how the testing culture that emerged in (product-controlled) American education systems in various ways influenced European countries with a long-standing (process-controlled) examination culture. These developments are analysed using the framework of the three transnational trends of educational assessment and exemplified with the cases of Sweden and Norway,

58 Brookhart, "The Use of Teacher Judgement," 71.

59 Race (Carson, 2008) and gender (Lundahl, 2006) issues are important to understanding how psychometric testing became popular; however, this chapter limits its focus on the interrelationship between premises of governing education systems and procedures for certifying and selecting individual students, and how (the magnitude and modes of) participation in transnational research projects may have prompted or reinforced these developments.

which responded in different ways to the (largely American-led) transnational meritocracy and accountability trends.

Transnational Trends Shaping the Assessment Cultures of Sweden and Norway

This section demonstrates that product-controlled education systems are more inclined to be receptive to the accountability trend than process-controlled education systems. When transnational emphasis on accountability increased throughout the twentieth century, these national states already had education systems that where built for independent measures of outcomes. Process controlled education systems, on the other hand, relies on meritocratic procedures that constitutes the teachers' authority (and licence them) to make judgments. Process controlled education systems did not have the capacity to embed the accountability demands in their existing procedures for determining merit. Thus, in these countries one can observe a separation between national assessment instruments used for meritocratic and accountability purposes, whereas in product controlled education system these may be included in the same national assessment instrument.

Sweden Adopts Psychometric Tests in Concert with the Transnational Meritocracy Trend

Norway and Sweden have a long-standing tradition of using national examinations to distinguish between students' levels of attainment. Both countries were later exposed to the progressive movement, where psychometric tests were perceived as an important tool for identifying students of special needs.[60] The use of psychometric tests in general education were, however, received in different ways by the public and the teacher profession in the two countries.

Swedish educators were highly involved in the American-led development of new psychometric instruments after World War II.[61] The Swedish researcher Frits

60 Lundahl, "Viljan att veta vad andra vet: Kunskapsbedömning i tidigmodern, modern och senmodern skola" (PhD diss., Uppsala University, Sweden, 2006); Jarning and Aas, "Between Common Schooling and the Academe"; Christian Ydesen, Kari Ludvigsen and Christian Lundahl, "Creating an Educational Testing Profession in Norway, Sweden and Denmark, 1910–1960," *European Educational Research Journal* 12, no. 1 (2013), doi: https://doi.org/10.2304/eerj.2013.12.1.120.

61 Christian Lundahl and Daniel Pettersson, "Den svenska skolens resultat. Från standardprov til PISA," in *Pisa – sannheten om skolen?*, ed. Eyvind Elstad and Kirsten Sivesind (Oslo: Universitetsforlaget, 2010); Florian Waldow, "Undeclared Imports:

Wigforss' contribution to the IEI study formed the beginning of a series of studies identifying low predictive validity of the Swedish examinations. The IEI project offered a basis for criticising the current examination system. Reporting on the influence of the Swedish contribution to the IEI project, Lundahl notes that Wigforss far from implemented 'American tests' in the Swedish education system.[62] Instead — drawing on the psychometric competence he had access to through the IEI project and beyond — Wigforss pushed for a Swedish 'twist' to the use of psychometric tests. Wigforss was convinced that teachers, when equipped with sufficient standardized instruments, were more capable of making comparable judgments than the existing examination system. Wigforss was at the time also involved in a governmental report which was investigating prospects of abolishing the examination entrance tests and instead let elementary school marks serve as instruments for selection. "If standardized marks could show better correlation with school success, then entrance tests would be unnecessary."[63]

As such, the IEI study marked the beginning of a blend of the examination and testing cultures, with larger emphasis given to American approaches to psychometric testing, albeit as a basis for helping teachers taking an even larger responsibility for certifying students' learning. The utilisation of psychometric tests was believed to provide more comparable measures and thus gave legitimacy to a transition where teachers were given more responsibility for grading based on tests developed through psychometric scientific principles. The psychometric expertise in Sweden emerged under large influence from American scholars in the IEI study and beyond, in particular through the State Psychological and Pedagogical Institute (SPPI) that was established in 1942. Lundahl observes that the participation in the IEI study helped Swedish researchers and policymakers to allege the need of a modern institution bringing science and educational practice closer together.[64] SPPI was established to develop psychometric competence in Sweden, and with a specific notion that "one important task for the Institute should be to develop new forms of tests that could substitute the entrance tests."[65]

Silent Borrowing in Educational Policy-making and Research in Sweden," *Comparative Education* 45, no. 4 (2009), URL: http://www.jstor.org/stable/40593191.

62 Christian Lundahl, "Inter/National Assessments as National Curriculum: The Case of Sweden," in *An Atlantic Crossing? The Work of the International Examination Inquiry, its Researchers, Methods and Influence* (London, UK: Symposium Books, 2008).

63 Ibid., 160.

64 Ibid.

65 Ibid., 172.

In the subsequent decades SPPI was a key institution for the termination of the Swedish entrance examinations as part of the reform in the 1960s, when Sweden unified its parallel school system to comprehensive schools.[66] Lundahl and Waldow observe that SPPI played an important role "as a producer and mediator of a standardized language; connecting diverging interests and creating the techniques to sustain an individualised and meritocratic education."[67]

Torsten Husén, professor at the Stockholm Teacher College from 1956 to 1971, was a prominent scholar in Sweden who exercised large influence on the education system for decades. As the chair of the IEA from 1962 to 1979, during which time it embarked on several studies in mathematics and science, Husén was in the position to project new global standards for educational assessment on Sweden's meritocratic procedures.[68] His recognition in Sweden was partly a product of the large international recognition he had being part of international research projects. Sweden participated in all eight IEA studies from 1960 to 1970, including the "Pilot Twelve-Country study" (1960), the "First International Mathematics Study" (1964) and the "First International Science Study" (1970–71) (Appendix 1). During this decade, the Swedish national tests (*standardprov*) gained preference over the traditional examinations, which were ultimately terminated in 1968.

The incremental transition from examinations to test-based certification procedures from the 1930s to the 1960s can be related to two wider features of Swedish society: the emphasis on psychological theories and methods, and the centralised governing tradition during this period. One may argue that Wigforss and Husén's contributions to IEI and IEA, respectively, reinforced these distinct features of Swedish society. Through these collaborative efforts, American psychometric theories and methods for testing student attainment made 'an Atlantic crossing'[69] and were incorporated into the education system in a distinct Swedish fashion. Tests were perceived as principle tools assisting teachers in making comparable judgments and as such they in some respects initially represented a strengthening of teachers' autonomy compared to the previous examination system in which the

66 Bo Lindensjö and Ulf P. Lundgren, *Utbildningsreformer och politisk styrning* (Stockholm: Liber, 2014).

67 Christian Lundahl and Florian Waldow, "Standardisation and 'Quick Languages': The Shape-Shifting of Standardized Measurement of Pupil Achievement in Sweden and Germany," *Comparative Education* 45, no. 3 (2009): 368, doi: 10.1080/03050060903184940.

68 Lundahl, "Inter/National Assessments."

69 Lawn, *An Atlantic Crossing*.

state — through higher education institutions — controlled teachers.[70] While the implications of the IEI and IEA studies varied across participating countries, these collaborative efforts can be viewed as milestones in what Lundahl and Waldow describe as the *first cycle of standardisation* in European education systems.[71]

In the 1990s, Sweden embarked on a more decentralised and market-based organisation of the education system.[72] While the national tests also had a role in the governing of the education system at the time of implementation in the 1960s, they became more important tools for controlling the more output-oriented and product-controlled education system that emerged from the 1990s. As the national tests were already based on psychometric principles that increasingly gained preference as a basis for governing education systems, it was not necessary to implement a new testing programme in response to the transnational *accountability* trend. Unlike Norway (discussed below), Sweden could simply expand its existing testing programme.

Recently the Swedish National Agency for Education also put emphasis[73] on formative use of the national tests. Its official webpage expresses certification and governing as key purposes, yet adds that "the national tests can also help specify curricula and subjects plans, and improve student achievement."[74] The recent addition of the emphasis on *supporting* learning and instruction can be perceived in view of the transnational *Assessment for learning* trend, that help the authorities to legitimise the expansion of the testing programme.[75] As such, since it was implemented primarily for *certification* from the 1930s through the 1960s, the national testing programme has accumulated the roles of *governing* and *supporting* learning and instruction in concert with the associated transnational *meritocracy, accountability* and *Assessment for learning* trends respectively.

Norway Adopts Psychometric Tests in Concert with the Transnational Accountability Trend

Norway did not have prominent contributors to the IEI study and IEA, as Sweden did. Norway's relatively limited contribution to the IEI study can be observed

70 Lundahl and Tveit, "Att legitimera nationella prov."
71 Lundahl and Waldow, "Standardisation and 'Quick Languages'."
72 Johanna Ringarp, *Professionens problematik* (Stockholm: Makadam, 2011).
73 Swedish National Agency for Education, "*National tests*," accessed June 30, 2015, http://www.skolverket.se/bedomning/nationella-prov.
74 Ibid.
75 Sverre Tveit, "Ambitious and Ambiguous."

when comparing the Swedish and Norwegian delegations' project reports.[76] Furthermore, reporting of the IEI study and further collaboration were constrained by the German occupation (1940–45) during World War II.[77] While Swedish members in the research team belonged to the progressive movement in primary education, Norwegian members of the research team were based in higher education and affiliated with secondary education.[78] As opposed to their fellow Swedish IEI members, they were not in a good position to influence the use of assessments in general education. The main institution where they could exercise influence was the University of Oslo, where the Department of Educational Research (*Pedagogisk forskningsinstitutt*) had been established in 1938. The head of the department, Johs Sandven — 'Norway's Husén' — had been visiting Edward Thorndike and colleagues at Teachers College in New York and was committed to developing psychometric tests for use in general education in Norway. He was, however, not as successful in establishing an institutional environment of psychometric testing as his Swedish counterpart. In the late 1960s controversies over the establishment of educational measurement as an academic discipline occurred in concert with the democratisation and increased student influence on university policies.[79] Sandven had to step down in 1972.

As shown in Appendix 1, unlike Sweden, Norway did not participate in the IEA studies until the Second International Science Study (1983–84) and the Reading Literacy Study (1990–91). Therefore, despite Norway modelled its education system on its Swedish neighbour after World War II,[80] weaker engagement with (and less implications of) the IEI and IEA studies from the 1930s to the 1960s may be one of the explanations as to why Norway did not follow Sweden in tak-

76 Archive Observations, the Carnegie Collections, the Rare Book and Manuscript Library, Columbia University, April 15th 2016. International Examinations Enquiry Committee, Norway, 1929–1937; International Examinations Enquiry Committee, Sweden, 1929–1937.

77 Kay Piene, *Eksamenskarakterer og forhåndskarakterer* (Oslo: Cappelen, 1961); Jarning and Aas, "Between Common Schooling."

78 Jarning and Aas, "Between Common Schooling."

79 Kim G. Helsvig, "Pedagogikkens grenser. Kampen om norsk pedagogikk ved Pedagogisk forskningsinstitutt 1938–1980" (Oslo: Abstract forlag, 2005).

80 Francis Sejersted, *Sosialdemokratiets tidsalder. Norge og Sverige i det 20.* århundre (Oslo: Pax forlag, 2005); Alfred Telhaug Oftedal, Odd Asbjørn *Mediås* and Petter Aasen, "From Collectivism to Individualism? Education as Nation Building in a Scandinavian Perspective." *Scandinavian Journal of Educational Research* 48, no. 2 (2004), doi: 10.1080/0031383042000198558.

ing on psychometric approaches to educational assessment in general education in the 1950s and 1960s.

At the time Norway attempted to introduce psychometric testing, in the late 1960s, resistance towards American psychometric approaches to determining merit flourished in Scandinavia. Seen from a Scandinavian progressivist educator of the 1960's perspective, it was too late for the Swedes to reject American standardized tests, while in Norway there was still time. Due to protests from the teacher profession and left-wing intellectuals, initial attempts to implement national testing in Norway in 1968 failed.[81]

It would take another three decades until standardized tests were implemented in full in Norwegian schools. As part of what Lundahl and Waldow calls the *second cycle of standardisation*, which I have called the *transnational accountability trend*, emphasis was put on holding schools and teachers accountable for student outcomes.[82] Norway's outcomes on the TIMSS 1995 study had raised some concerns, but it did not cause the same public outcry that followed the publication of the first PISA tests in 2001. What is often labelled the 'PISA shock' prompted several European countries that had taken a reluctant attitude towards standardized testing (Denmark and Germany are other notable examples) to implement new national testing programmes.[83]

In Norway an OECD report of 1988 had expressed criticism for the country's lack of a system for monitoring student performance as a way to hold municipalities accountable for learning outcomes.[84] Government committees and the Ministry and Parliament discussed a system for national evaluation of schooling throughout the 1990s.[85] However, it was not until after the PISA shock that a national system for quality assessment was implemented. The PISA and OECD influence is illustrated well by the words of the Norwegian minister of education,

81 Forsøksrådet for skoleverket, *Standardiserte prøver i skolen. Forsøk og reform i skolen – Nr. 16* (Oslo: Universitetsforlaget, 1969).

82 Lundahl and Waldow, "Standardisation and 'Quick Languages'."

83 Aaron Benavot and Erin Tanner, "The Growth of National Learning Assessments in the World, 1995–2006" (Background paper for the EFA global monitoring report: Education For All by 2015: Will We Make It?. Paris: UNESCO, 2007).

84 OECD, *OECD-vurdering av norsk utdanningspolitikk* (Oslo: Kirke- og undervisningsdepartementet, 1988).

85 Marit K. Granheim, Ulf P. Lundgren and Tom Tiller, *Målstyring og evaluering i norsk skole. Sluttrapport EMIL-prosjektet, NORAS/LOS,* LOS-notat nr. 7 (Oslo: Norges råd for anvendt samfunnsforskning, 1990); OECD, *OECD-vurdering av norsk utdanningspolitikk; Stortingsmelding nr. 47*; *Stortingsmelding nr. 28.*

who, due to the disappointing results of the first PISA tests, stated that it was "almost like coming home from a winter Olympics without one Norwegian medal."[86] This power of the league tables that the PISA studies[87] produce can be related to the tests' close connection with the role of the OECD as a global policy agency[88] and how this information is used by governments to legitimise reforms.[89] A new national testing program was implemented in Norway in 2004, initially motivated by the need for information to be used in the governing of education.[90]

Table 3: The National Assessment Instruments in Sweden and Norway

Country	**SWEDEN**	**NORWAY**	
Instrument	**National tests (prov)**	**National examinations**	**National tests (prøver)**
Year and subject/skill	Year 3: Mathematics, Swedish, and Swedish as a second language. Year 6: Mathematics, Swedish, Swedish as a second language, and English. Year 9: Mathematics, Swedish, Swedish as a second language, and English. Additionally, one natural science-oriented test (biology, physics, or chemistry) and one social science-oriented test (geography, history, religion, or social science).	Year 10: One examination in either Norwegian, English, or mathematics	Year 5: English, reading, and numeracy. Year 8: English, reading, and numeracy. Year 9: Reading and numeracy.

86 Helge O. Bergesen, *Kampen om Kunnskapsskolen* (Oslo: Universitetsforlaget, 2006), 41.
87 Meyer and Benavot, "Introduction."
88 Daniel Pettersson, "Internationell kunskapsbedömning som inslag i nationell styrning av skolan" (PhD diss., Uppsala University, 2008).
89 Sverre Tveit, "Educational Assessment in Norway," *Assessment in Education: Principles, Policy & Practice* 21, no. 2 (2014), doi: 10.1080/0969594X.2013.830079.
90 Tveit, "Educational Assessment in Norway."

Country	SWEDEN	NORWAY	
Instrument	**National tests (prov)**	**National examinations**	**National tests (prøver)**
Instrument developer	Developed by expert groups at universities commissioned by the Swedish National Agency for Education.	Developed by expert groups of teachers and scholars, commissioned by the Norwegian Directorate for Education and Training	Developed by expert groups at universities, commissioned by the Norwegian Directorate for Education and Training.
Marking procedures	The teachers mark the responses themselves, based on guidelines from the Swedish National Agency for Education	Two external and trained examiners mark the responses based on guidelines from the Norwegian Directorate for Education and Training.	Auto-computerised marking of most items. For open questions in the reading tests, teachers assign scores, based on guidelines from the Norwegian Directorate for Education and Training.
Subject or skill orientation	The instruments are constructed based on disciplinary goals stated in the curriculum for the respective subjects.	The instruments are constructed based on disciplinary goals (competence aims) stated in the curriculum for the respective subjects.	The instruments are constructed based on the basic skills, which are integrated in the competence aims for all subjects' curriculum.
Implemented	1930–1960s	Emerged in the 1800s	2004
Certification role	1960s	1800s	
Governing role	1960s (increased from the 1990s)	1800s	2004
Support role	2004		2006

Due to alleged overemphasis on school accountability, including student boycotts that jeopardised the validity of the assessment data,[91] the policy discourse changed in 2005. A one-year moratorium was held due to substantial problems

91 Halvard Hølleland, "Nasjonale prøver og kvalitetsutvikling i skolen," in *Elevvurdering i skolen – grunnlag for kulturendring*, ed. Sverre Tveit (Oslo: Universitetsforlaget, 2007).

with the testing programme and controversies over the publication of league tables,[92] and, following a change of government, radical changes were undertaken to ensure the legitimacy of the testing programme. Upper secondary education tests were terminated, and compulsory education tests were moved from the conclusion of Year 4 and 7 to the beginning of the subsequent years (Year 5 and Year 8). This change reflected a shift of purposes where new emphasis was put on the tests' role in supporting, along with governing, learning and instruction.

This change can be interpreted in view of the increased transnational emphasis on the *Assessment for learning*, and resistance to the *accountability* trend that was perceived to have dominated the new education reform and associated national testing programme. Despite less emphasis on the publication of league tables, *governing* remained the key purpose of the national tests, although the government and its executive agency also stressed their role in *supporting* learning and instruction.

Concluding Discussion: National Assessment Instruments' Accumulation of Roles in Concert with Transnational Trends of Educational Assessment

In this chapter I have demonstrated that different premises of process- and product-control partly explain the emergence of two distinctly different approaches to educational assessment in the continental European countries and the United States: The emphasis on professional (subjective) judgments and external (objective) measurement respectively. I have explained how both the increased emphasis on the *certification* and *governing* roles of educational assessment prompted the psychometric testing technology in the United States, and I further addressed how other countries took up the potential of using national tests for these purposes through the transnational *meritocracy* and *accountability* trends in the second half of the twentieth century. With the examples of Sweden and Norway, I have demonstrated how countries that began with an examination culture developed in different directions, which is, in part, related to level of engagement with the transnational *meritocracy* and *accountability* trends throughout the twentieth century.

Whereas the *meritocracy* trend from the 1930s brought psychometric approaches to educational assessment to Sweden (and ultimately the replacement of examinations with psychometric tests by the 1960s), it was through the *accountability* trend

92 Eyvind Elstad, "Schools Which Are Named, Shamed and Blamed by the Media: School Accountability in Norway," *Educational Assessment, Evaluation and Accountability* 21, no. 2 (2009), doi:10.1007/s11092009-9076-0; Svein Lie et al., *Nasjonale prøver på ny prøve* (Oslo: Department of Teacher Education and School Research, University of Oslo, 2005).

from the 1990s onwards that psychometric approaches to educational assessment broke through in Norway. As shown in Table 3, the result of this is that Norway currently has two national assessment instruments, one with a *certification* role and one with a *governing* role. When the *accountability* trend brought increased emphasis on psychometric testing, Sweden could instead simply strengthen its existing national testing programme. At the turn of the millennium, the transnational emphasis on *Assessment for learning* contributed to a new emphasis on assessment instruments' role in *supporting* learning and instruction. Both countries have 'added' this purpose to its respective national tests. As such, contemporary uses of national examinations and tests should be understood in view of the accumulation of educational assessment roles in concert with the transnational emphasis on *meritocracy, accountability* and, *Assessment for learning*.

It is essential to acknowledge the different timings of the implementation of national tests if different cultures of educational assessment are to be understood. As demonstrated in Table 3, the national tests in Sweden are subject and disciplinary based in accordance with the IEI research project that shaped the meritocracy trend. This reflects how they are used to *certify* (subject) learning. In Norway the national tests are interdisciplinary and skills based, which reflects the emphasis on skills in PISA, the most influential comparative testing programme associated with the *accountability* trend.

Thus, while they are called 'national tests' in both Norway and Sweden, these assessment instruments underwent completely different transnational influences characteristic to the transnational trend at the time of implementation. These differences may be illustrative of wider patterns in European countries' cultures of educational assessment. Similar to Norway, Denmark and Germany opposed implementation of psychometric tests during the *meritocracy* trend that emerged from the 1930s. In these countries such tests did not break through until the *accountability* trend that took firm root in the 1990s and was further propelled by PISA shocks at the turn of the millennium. Other countries' developments may have been more similar to that of Sweden, which replaced its existing national examination programme. Being already based on psychometric principles, only strengthening and expansion of existing testing programmes were needed to respond to the *accountability* trend.

Conclusively, as observed in both Norway and Sweden, all countries are likely to be affected by the *Assessment for learning* trend. The recent emphasis on formative assessment can both be associated with a genuine change of focus from 'summative' to 'formative' assessment, as advocated by many scholars and policymakers. It may however also reflect a legitimation strategy intended to make sure that teachers accept national tests in which *governing* learning and instruction remains the

principal purpose albeit in a more attractive wrapping. The promulgation of the three roles and the three transnational trends of educational assessment undertaken in this chapter help envision how national assessment instruments have come to accumulate multiple purposes in response to different transnational developments at the time of implementation, revision and legitimation.

Literature

Aas, Gro H. "Likhet uten solidaritet? Idéhistoriske studier av karakterer I utdanning og meritokrati." PhD diss., University of Gothenburg, 2006.

Ball, Stephen J. "The Teacher's Soul and the Terrors of Performativity." *Journal of Education Policy* 18, no. 2 (2003): 215–28. doi:10.1080/0268093022000043065.

Benavot, Aaron, and *Erin Tanner*. "The Growth of National Learning Assessments in the World, 1995–2006." Background paper for the EFA global monitoring report: Education For All by 2015: Will We Make It? Paris: UNESCO, 2007.

Bergesen, Helge O. Kampen om Kunnskapsskolen. Oslo: Universitetsforlaget, 2006.

Black, Paul, and *Dylan Wiliam*. "Assessment and Classroom Learning." *Assessment in Education* 5, no. 1 (1998): 7–74. doi:10.1080/0969595980050102.

Brookhart, Susan M. "The Use of Teacher Judgement for Summative Assessment in the USA." *Assessment in Education: Principles, Policy & Practice* 20, no. 1 (January 30, 2013): 69–90. doi:10.1080/0969594X.2012.703170.

Carson, John. *The Measure of Merit: Talents Intelligence, and Inequality in the French and American Republics, 1750–1940*. Princeton, NJ: Princeton University Press, 2007.

Dale, Roger. "Globalisation, Knowledge Economy and Comparative Education." *Comparative Education* 41, no. 2 (2005): 117–49. doi:10.1080/03050060500150906.

Elstad, Eyvind. "Schools Which Are Named, Shamed and Blamed by the Media: School Accountability in Norway." *Educational Assessment, Evaluation and Accountability* 21, no. 2 (2009): 173–89. doi:10.1007/s11092-009-9076-0.

Forsberg, Eva. "Utbildningens Bedömningskulturer i Granskningens Tidevarv." *Utbildning & Demokrati* 23, no. 3 (2014): 53–76.

Forsøksrådet for skoleverket. *Standardiserte prøver i skolen. Forsøk og reform i skolen – nr 16*. Oslo: Universitetsforlaget, 1969.

Granheim, Marit K., Ulf P. Lundgren, and *Tom Tiller*. *Målstyring og evaluering i norsk skole. Sluttrapport EMIL-prosjektet NORAS/LOS-i-utdanning*. LOS-notat nr. 7. Oslo: Norges råd for anvendt samfunnsforskning, 1990.

Grek, Sotiria, *Martin Lawn*, *Bob Lingard*, *Jenny Ozga*, and *Risto Rinne*. "National Policy Brokering and the Construction of the European Education Space in

England, Sweden, Finland and Scotland." *Comparative Education* 45, no. 1 (2009): 5–21. doi:10.1080/03050060802661378.

Harlen, Wynne. "On the Relationship Between Assessment for Formative and Summative Purposes." In *Assessment and Learning*, edited by John Gardner, 87–102. London, UK: Sage, 2006. doi: 10.4135/9781446250808.n6.

Hattie, John. *Visible Learning*. London, UK: Routledge, 2008.

Hayward, E. Louise. "Curriculum, Pedagogies and Assessment in Scotland: The Quest for Social Justice. 'Ah Kent Yir Faither'." *Assessment in Education: Principles, Policy & Practice* 14, no. 2 (July 2007): 251–68. doi:10.1080/09695940701480178.

Helsvig, Kim G. Pedagogikkens grenser. Kampen om norsk pedagogikk ved Pedagogisk forskningsinstitutt 1938–1980. Oslo: Abstract forlag, 2005.

Hopfenbeck, Therese N., *Maria Teresa Flórez Petour*, and *Astrid Tolo*. "Balancing Tensions in Educational Policy Reforms: Large-Scale Implementation of Assessment for Learning in Norway." *Assessment in Education: Principles* 22, no. 1 (2015): 44–60.

Hopfenbeck, Therese N., *Astrid Tolo*, *Maria Teresa Florez*, and *Yasmine El Masri*. "Balancing Trust and Accountability? the Assessment for Learning Programme in Norway." 2013. Accessed May 21t, 2015. http://www.oecd.org/edu/ceri/Norwegian%20GCES%20case%20study%20OECD.pdf.

Hopmann, Stefan T. "On the Evaluation of Curriculum Reforms." *Journal of Curriculum Studies* 35, no. 4 (2003): 459–478. doi: 10.1080/00220270305520.

Hølleland, Halvard. "Nasjonale prøver og kvalitetsutvikling i skolen." In *Elevvurdering i skolen – grunnlag for kulturendring*, edited by S. Tveit, 29–44. Oslo: Universitetsforlaget, 2007.

IEA. "Completed Studies." *IEA*. Accessed May 30, 2015. http://www.iea.nl/completed_studies.html.

Jarning, Harald, and *Gro H. Aas*. "Between Common Schooling and the Academe: The International Examinations Inquiry in Norway, 1935–1961." In *An Atlantic Crossing? The Work of the International Examination Inquiry, its Researchers, Methods and Influence*, 181–204, edited by M. Lawn. Oxford, UK: Symposium Books, 2008.

Jonsson, Anders, *Christian Lundahl*, and *Anders Holmgren*. "Evaluating a Large-Scale Implementation of Assessment for Learning in Sweden." *Assessment in Education: Principles, Policy & Practice* 22, no. 1 (2015): 104–21. doi:10.1080/0969594X.2014.970612.

Kamens, David H. "Globalisation and the Emergence of an Audit Culture: PISA and the Search for 'Best Practices' and Magic Bullets." In *PISA, Power, and Policy: The Emergence of Global Educational Governance*, edited by Heinz-Dieter Meyer, and Aaron Benavot, 117–139. Oxford, UK: Symposium Books, 2015.

Kellaghan, Thomas. "The Globalisation of Assessment in the 20th Century." *Assessment in Education: Principles* 8, no. 1 (2001): 87–102. doi: 10.1080/0969594012003327O.

Lawn, Martin. ed. *An Atlantic Crossing? The Work of the International Examination Inquiry, its Researchers, Methods and Influence*. London, UK: Symposium Books, 2008.

Lawn, Martin. "Voyages of Measurement in Education in the Twentieth Century: Experts, Tools and Centres." *European Educational Research Journal* 12, no. 1 (2013): 108–119. doi:10.2304/eerj.2013.12.1.108.

Lie, Svein, *Therese N. Hopfenbeck*, *Elisabeth Ibsen*, and *Are Turmo*. *Nasjonale prøver på ny prøve*. Oslo: Department of Teacher Education and School Research, University of Oslo, 2005.

Lindensjö, Bo, and *Lundgren, Ulf P*. *Utbildningsreformer och politisk styrning*. Stockholm: Liber, 2014.

Lundahl, Christian. "*Viljan att veta vad andra vet: Kunskapsbedömning i tidigmodern, modern och senmodern skola*." PhD diss., Uppsala University, Sweden, 2006.

–. "Inter/National Assessments as National Curriculum: The Case of Sweden." In *An Atlantic Crossing? The Work of the International Examination Inquiry, its Researchers, Methods and Influence*, edited by Martin Lawn, 157–180. Oxford, UK: Symposium Books, 2008.

Lundahl, Christian, and *Daniel Petterson*. "Den svenska skolens resultat. Från standardprov til PISA, in *PISA: sannheten om skolen?*, edited by Eyvind Elstad and Kirsten Sivesind, 222–239. Oslo: Universitetsforlaget, 2010.

Lundahl, Christian, and *Florian Waldow*. "Standardisation and 'Quick Languages': The Shape-Shifting of Standardized Measurement of Pupil Achievement in Sweden and Germany." *Comparative Education* 45, no. 3 (2009): 365–385. doi: 10.1080/03050060903184940.

Madhaus, George F. "Ralph Tyler's Contribution to Program Evaluation." In *Evaluation Roots. A Wider Perspective of Theorists' Views and Influences*, 2nd edition, edited by Marvin C. Alkin, 157–164. Los Angeles, CA: Sage, 2013.

Meyer, Heinz-Dieter, and *Aaron Benavot*. "Introduction." In: *PISA, Power, and Policy: The Emergence of Global Educational Governance*, edited by Heinz-Dieter Meyer and A. Benavot, 9–26. Oxford, UK: Symposium Books, 2015.

Müller Detlef, *Fritz K. Ringer*, and *Brian Simon*. *The Rise of the Modern Educational System*. Cambridge, UK: Cambridge University Press, 1986.

Nordin, Andreas, and *Daniel Sundberg*. "The Making and Governing of Knowledge in the Education Policy Field." In *Transnational Policy Flows in European Education*, edited by Andreas Nordin, and Daniel Sundberg, 9–20. Oxford, UK: Symposium Books, 2014.

Organisation for Economic Co-operation and Development [OECD]. OECD-vurdering av norsk utdanningspolitikk. Oslo: Kirke-og undervisningsdepartementet, 1988.

–. *Formative Assessment: Improving Learning in Secondary Classrooms*. OECD, 2005.

–. "Centre for Educational Research and Innovation (CERI) – What Works." Accessed July 21, 2015. https://www.oecd.org/edu/ceri/centreforeducationalresearchandinnovationceri-whatworks.htm.

–. *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*, OECD, 2013. Accessed July 22, 2015. http://www.oecd.org/edu/school/oecdreviewonevaluationandassessmentframeworksforimprovingschooloutcomescountryreviews.htm.

Pettersson, Daniel. "Internationell kunskapsbedömning som inslag i nationell styrning av skolan." PhD diss., Uppsala: Uppsala University, 2008.

Piene, Kay. *Eksamenskarakterer og forhåndskarakterer*. Oslo: Cappelen, 1961.

Prøitz, Tine S. "Uploading, Downloading and Uploading Again – Concepts for Policy Integration in Education Research." *Nordic Journal of Studies in Educational Policy* 1 (2015): 70–80. doi:10.3402/nstep.v1.27015.

Ringarp, Johanna. *Professionens problematic.* Stockholm: Makadam, 2011.

Sejersted, Francis. *Sosialdemokratiets tidsalder. Norge og Sverige i det 20. århundre*. Oslo: Pax forlag, 2005.

Swedish National Agency for Education. *National tests*. Accessed June 30, 2015. http://www.skolverket.se/bedomning/nationella-prov.

Telhaug, Alfred Oftedal, Odd Asbjørn Mediås, and *Petter Aasen*. "From Collectivism to Individualism? Education as Nation Building in a Scandinavian Perspective." *Scandinavian Journal of Educational Research* 48, no. 2 (2004): 141–158. doi: 10.1080/0031383042000198558.

Tröhler, Daniel. "Truffle Pigs, Research Questions, and Histories of Education." In *Rethinking the History of Education: Transnational Perspectives on Its Questions, Methods, and Knowledge*, edited by Thomas S. Popkewitz, 75–92. New York, NY: Palgrave Macmillan US, 2013.

Tveit, Sverre. "Ambitious and Ambiguous: Shifting Purposes of National Testing in the Legitimation of Assessment Policies in Norway and Sweden (2000–2017)." *Assessment in Education: Principles, Policy & Practice* (forthcoming, 2018).

Tveit, Sverre and Christian Lundahl. "New Modes of Policy Legitimation in Education: (Mis)using Comparative Data to Effectuate Assessment Reform". *European Educational Research Journal* (2017). doi: https://doi.org/10.1177/1474904117728846»

Tveit, Sverre. "Educational Assessment in Norway." *Assessment in Education: Principles, Policy & Practice* 21 no. 2 (2014): 221–237.

UNESCO. "World Declaration on Education for All: Meeting basic needs." Adopted by the World Conference on Education for All. New York, NY: UNESCO, 1990.

Ydesen, Christian, *Kari Ludvigsen*, and *Christian Lundahl*. "Creating an Educational Testing Profession in Norway, Sweden and Denmark, 1910–1960." *European Educational Research Journal* 12, no. 1 (2013): 120. doi: https://doi.org/10.2304/eerj.2013.12.1.120.

Young, Michael. *The Rise of Meritocracy*. London, UK: Penguin Books, 1958.

Waldow, Florian. "Undeclared Imports: Silent Borrowing in Educational Policy-making and Research in Sweden." *Comparative Education* 45, no. 4 (2009): 477–494. URL: http://www.jstor.org/stable/40593191.

Appendix 1: Participation in IEA and PISA studies[93]

Year	Test	Content	Age	Provider	NO	SE
1960	Pilot Twelve-Country Study	Mathematics, reading comprehension, geography, science, and non-verbal ability	13	IEA	-	X
1964	FIMS (First International Mathematics Study)	Mathematics	13	IEA	-	X
1970–71	FISS (First International Science Study)	Science	10, 14, final SE	IEA	-	X
1970–71	Six Subject Survey: Reading comprehension	Reading	10, 14, final SE	IEA	-	X
1970–71	Six Subject Survey: Literature Education	Literature	14, final SE	IEA	-	X
1970–71	Six Subject Survey: English as a Foreign language	English as a foreign language	14, final SE	IEA	-	X
1970–71	Six Subject Survey: French as a Foreign language	French as a foreign language	14, final SE	IEA	-	X

93 Information gathered from http://www.iea.nl/brief_history.html on April 17th 2016. Not included: Classroom Environment Study (1981–83); Computers in Education Study (COMPED) (1989, 1992); Preprimary project (PPP), 1987–89, 1992, 1995–97; Second Information Technology in Education Study Modul 1 (SITES-M1).

Year	Test	Content	Age	Provider	NO	SE
1970–71	Six Subject Survey: Civic Education	Civic Education	10, 14, final SE	IEA	-	X
1980–82	SIMS (Second International Mathematics Study)	Mathematics	13	IEA	-	X
1983–84	SISS (Second International Science Study)	Science	10, 14, final SE	IEA	X	X
1984–85	Written Composition Study	Writing	10–12; 15–17; 17–19	IEA	-	X
1990–91	Reading Literacy Study	Reading	9, 14	IEA	X	X
1994–95	TIMSS (The Third International Mathematics and Science Study)	Mathematics and Science	9, 13, final SE,	IEA	X	X
1995	Language Education Study	English, French, German, and Spanish.	15–16; 17–18	IEA	X	X
1998–99	TIMSS 1999	Mathematics and Science	Grade 8	IEA	X	X
2000	PISA 2000	Reading, Mathematics, Science	15	OECD	X	X
2001	PIRLS 2001	Reading	Grade 4	IEA	X	X
2003	TIMSS 2003	Mathematics and Science	Grade 4, 8	IEA	X	X
2003	PISA 2003	Reading, Mathematics, Science	15	OECD	X	X
2006	PIRLS 2006	Reading	Grade 4	IEA	X	X
2006	PISA 2006	Reading, Mathematics, Science	15	OECD	X	X
2007	TIMSS 2007	Mathematics and Science	Grade 4, 8	IEA	X	X
2009	PISA 2009	Reading, Mathematics, Science	15	OECD	X	X
2011	TIMSS 2011	Mathematics and Science	Grade 4, 8	IEA	X	X
2011	PIRLS 2011	Reading	Grade 4, 8	IEA	X	X
2012	PISA 2012	Reading, Mathematics, Science	15	OECD	X	X
2015	PISA 2015					X