Assessment and Evaluation

Authors:

Associate Professor Nelli Piattoeva (corresponding author; nelli.piattoeva@tuni.fi), Tampere University, Faculty of Education and Culture

Professor Jaakko Kauko, Tampere University, Faculty of Education and Culture

Dr Hannele Pitkänen, Tampere University, Faculty of Education and Culture

Dr Tommi Wallenius, University of Helsinki, Faculty of Educational Sciences

Introduction

This chapter examines assessment and evaluation in education as a political and power-laden issue. It describes cases from around the globe to argue that the policies and practices of assessment and evaluation manifest a growing and increasingly central area of education policy, and a policy instrument that makes contemporary societies and their formal education systems governable. Overall the chapter offers an historical overview of the development of evaluation, monitoring and testing practices, considering different literatures in the field of policy approaches to assessment and evaluation, the influence of international organisations on these processes, and the contingency of the expressions and translations of assessment and evaluation policies in national contexts.

Our approach moreover draws on the tradition of critical research in sociology and politics of education, aligned with Ozga, Dahler-Larsen, Segerholm and Simola (2011) or Gunter Hall and Mills (2015), for instance. This approach entails analysing the relationship between assessment and evaluation and transformations in politics, governance and forms of the state, viewing assessment and evaluation as concrete manifestations of such changes as well as means of promoting them. In this way changing goals and means of assessment shed light on how the relationship between governed and governing is changing, and how societies overall are governed. This view of assessment and evaluation is further anchored in post-structural theorising on the relationship between power and knowledge based on a set of assumptions (e.g. Ball 2003; 1998; Foucault 1982; 2000; Rose & Miller 2010; Rose, O´Malley & Valverde 2006). These include the notion of governmentality that understands power not as something possessed, but entailing a practice exercised discursively and materially by different actors including those of the state and others (Rose et al 2006). Such practices offer a means of exercising power indirectly, at a distance (Rose & Miller 2010), by shaping and

leading, rather than constraining and sanctioning (e.g. Foucault 1982). It follows that instruments – including assessment and evaluation and their constitutive elements, such as data collection practices – are not neutral but performative (Ball 1998; 2003; Grimaldi 2020), as they do not merely describe but also shape realities according to their logic.

Contemporary expansion of assessment and evaluation is embedded in the ongoing rationalisation of societies (cf. Le Gales 2016) of which education is both a target and a vehicle. This process of further rationalisation impacts different types of societies globally, thus subsuming different political regimes and priorities, i.e., from democratic policies of accountability to neoliberal policies of marketisation and competition and authoritarian and control-oriented policies primarily seeking further surveillance and intrusion into the lives of individual and collective actors (cf. Kipnis 2008). Thus, it is important to go beyond perceiving assessment and evaluation as recent and specific to the rise of neoliberal governmentality (cf. Olssen & Peters 2005), and instead see them as malleable for different and evolving political concerns and ideological stances (see e.g. Grimaldi 2020; Kipnis 2008). Moreover, precisely because assessment and evaluation have longer albeit intricate historical roots, their contemporary proliferation is not uniform across countries, political regimes or institutional landscapes. Consequently, this chapter works through both historical and contemporary cases of assessment and evaluation to highlight and explore contextual specificity.

A review of evaluation research yields no simple or shared definition of assessment and evaluation. For the purposes of this chapter, we use the two terms synonymously and interchangeably (see Centeno, Kauko & Candido 2017 for umbrella concepts) and acknowledge that they may assume many forms (Kellaghan, Stufflebeam & Wingate 2003). They generally entail a phenomenon and a process of judging the worth and value of evaluated entities in relation to a predetermined standard or frame, aiming at demonstrating 'quality' to outsiders (Harvey 2004) and retrospective learning (Vedung 2010). Assessment varies in scale and scope, and which evaluative techniques and practices are put into practice (see e.g. Dahler-Larsen 2012a: 5–13). Importantly, evaluation changes the constitution of social relations and redefines social reality, as we already argued above (Dahler-Larsen 2012b; 2019).

Our primary concern is with school education and assessment as a policy area and a tool of policy. Given the vast array of evaluation tools, we limit our focus to instances where standardised assessments of student or teacher performance are used to pass judgment on aggregate education outcomes at national or global levels of education decision-making. By *policy area* we highlight that assessment and evaluation have grown into a central focus of education policy, both nationally and internationally (Verger, Parcerisa & Fontdevila 2019; Kauko, Takala & Rinne 2018a). In other words,

it is an issue that increasingly faces concerted attempts at definition, intervention and monitoring by legislative, executive and expert authorities. Assessment and evaluation have become both "normal" and "expected" action (Ozga et al 2011; Dahler-Larsen 2012b) and a means to strive for and demonstrate tangible policy results in the sphere of education and beyond (Kauko et al. 2018a). In addition, we view assessment and evaluation as *policy instruments*, that is, as practical devices of policy operationalisation and implementation, and as condensed (but contested) forms of knowledge about social control and ways of exercising it that structure policy according to their logic (e.g. Lascoumes & Le Gales 2007). These two intertwined conceptualizations enable us to offer rich perspectives on the political and power-laden nature of assessment.

From the perspective of evaluation as a policy instrument, it helps to further specify them as "data-intensive instruments" (cf. Grek, Maroy & Verger 2021) because their power and ability to govern at a distance emanate from and are rooted in the complex processes of data collection, publication and use. The prominent role of data in assessment and evaluation demonstrates the 'scientisation' of education policy and governance (Grek & Ozga 2010), as data produced with complex statistical expertise confers scientific authority upon both macro policy-making and micro decisions on education. In particular, the cross-national proliferation of national and international large-scale assessments – albeit publicised and used in different ways, as the chapter will show – enables different evaluation practices. These may include the application of data as seemingly apolitical evidence in policy decisions on curriculum, funding or individual remuneration of teachers or educational organisations, and the publication of assessment results in the media and thus for use outside of the traditional policy-making contexts such as governmental agencies. For this reason, we propose to examine assessment and evaluation as areas of policy and policy instruments through the heuristic lens of data and specifically the evolution of data practices. Looking at the evolution of data practices as a ubiquitous component of assessment policies enables 1) to capture both the policy area and policy instrument aspects of assessment and evaluation; 2) to understand the pervasive relation between power and knowledge; and 3) to focus attention on contextual factors that may be glossed over in both policy rhetoric and academic literature preoccupied with documenting global convergence.

The chapter proceeds as follows. The next section explores its historical evaluation through a number of cases that demonstrate how it has evolved into a tool of regulation and information-gathering by state authorities. As evaluation was scaled up to the national level it also, following progress in statistical and technical possibilities, became increasingly standardised, numerical, and large-scale. In this section, we also demonstrate standardised testing and school inspections as common formats of evaluation. The following section presents a description of contemporary developments in the

governance of societies, which we deem essential for understanding how and why evaluation has risen to greater prominence. This analysis is followed by empirical illustrations of assessment and evaluation on two intertwined scales of governance: global and national. These have become more entangled through an intensification of international and national standardised testing. We show how, on the one hand, there is a general consensus as to the purposes and means of implementing assessment and evaluation. At the same time, coexisting with this global convergence, there are national variants on both political and practical levels. In all these discussions, we pursue the political and power-laden nature of assessment and evaluation that we examine through the lens of evolving data practices.

Historicity and contextuality of evaluation and assessment

In the current boom of assessment and evaluation, it is easy to overlook their long and multi-layered histories that reveal their evolving political roles (see e.g. Alarcón & Lawn 2018; Saari 2013). In addition, evaluation appears to be contingent, and this contingency is partly rooted in specific local histories of assessment (Alarcon & Lawn 2018; Kauko et al 2018a; Verger et al. 2019; Suominen et al. 2018). Thus, contemporary intensification of evaluation nationally and globally cannot be attributed solely to the triumph of market-driven policy reforms, often labelled as neoliberal (see e.g. Olssen & Peters 2005). The current policies of educational assessment need to be acknowledged as a continuum, albeit far from linear or unquestioned, of multi-layered histories of envisioning and implementing education governance. The political role of assessment and the conditions that shape it can only be understood in relation to socio-historical trajectories including changing rationalities, discourses, policies and practices of evaluation (see e.g. Rose & Miller 2010; Grimaldi 2020), and also other contextual elements, such as changing political climate and interests, actors, value systems and power relations as well as evolving statistical and technological affordances. We illustrate this argument by examining diverse historical and geographical cases of school inspection and standardized testing as common techniques of evaluation.

The rise of school inspection as a mode of assessment and evaluation by the state

In Western European contexts, assessment in education has been practised at least since the Protestant Reformation in the sixteenth century (Saari 2011; Varjo, Simola & Rinne 2016). The earliest records of formal and regular assessment go back even further, to the Western Zhou Dynasty in China (1027–

1771 BC), where officials were selected for the imperial civil service through highly competitive examinations (Berry & Adamson 2011; Zhou 2019). In the Nordic region, since medieval times, bishops have monitored and reported on literacy and knowledge of catechisms in their bishoprics (Varjo et al. 2016). Thus, in Europe, the early roots of school inspection can be found in inspections mandated by religious authorities. For example, in Wales, before the formation of state organised inspection and mass education system, religious societies supervised the expenditure of money and ensured the teaching of religious tenets. (James & Davies 2009.)

Since the advent of modern mass public education in the latter half of the 19th century as part of the nation-building and state centralisation project in Europe, school inspection has shifted from church to secular inspectorates embedded in the governing bodies of the states (e.g. Varjo et al. 2016; Evertsson 2015). Inspectors, for instance, supported teachers in shifting from a narrow focus on religious content and literacy to a broader curriculum and new school subjects such as citizenship education (Evertsson 2015: 262), or reported and produced a statistical overview of public education for the administration and senate (Varjo et al. 2016: 23–26.). Thus, the data produced at the school and local levels were not only intended for use in specific local contexts, but also as "aggregate" data, that is, data that provide information on education at the national level for policymakers.

Inspection was also widely deployed for monitoring schools' and teachers' adherence to the norms and rules prescribed by laws and regulations. In Wales, the introduction of a government education grant of £20,000 in 1833 was followed by a requirement for quality in instruction and fiscal responsibility in the use of government money, which the state inspectorate should monitor. The established system was later followed by the initiation of "performance indicators for schools and teachers in the form of Payment by Results in 1862" together with a national inspectorate monitoring the achievement of the prescribed standards. (James & Davies 2009: 668.) The system made government grants to schools and teachers' salaries contingent upon the pass rates of students in reading, writing and arithmetic as well as attendance at school (Knudsen 2016: 510).

The development of standardised testing

Along with school inspection, standardised testing is a key mechanism in carrying out assessment. The idea of scientific standardised testing of pupil attainment and personal abilities and traits first emerged in the early 20th century in the United States, fostered by a positivist scientific paradigm and developments in educational psychology rooted in precise natural-science types of measurement of societal phenomena. These trends coincided with further progress in statistical thinking and

calculative practices, making it possible to not only focus on the educational attainment and abilities of individual pupils or schools, but also on aggregate pupil populations (Saari 2011; 2013). Technically, these developments enabled the emergence of statistical analyses of learning outcomes and comparisons between pupils and pupil populations. At the same time, these methods of data collection and analysis produced the idea of what learning is, or at least one version of it. The statistical understanding of normal distribution, the Gaussian curve, provided a visualisation method to imagine learning as a linear progression, and determine normal and deviant learning results within a sample of pupils, thus producing an idea of a "normal pupil" (Saari 2013; see also Hacking 1991).

Sweden has been a pioneering country in the production, availability and application of standardised testing data. As early as the 1940s, Sweden adopted standardised testing for pupil selection, but did not use testing to determine the performance of schools or teachers, contrary to the recent trend in many education systems. Rather, standardised testing was considered to provide an equal, transparent and commensurable basis for grading and selection for further education to be deployed by teachers only (Wallenius 2016).

In another historical context – that of the early Soviet Union – all standardized testing came under scrutiny in the 1930s and was suspended for more than forty years. Ranking individuals conflicted with socialist ideology. The Soviet authorities saw researchers engaged in testing as narrow-minded empiricists and their methods as tools for justifying inequality, with test results serving the purpose of early segregation and simplified accounts of learning abilities (Piattoeva & Gurova 2018). In addition, testing was viewed as "bourgeois" and by extension, a means of exploiting individuals, whereas the official ideology postulated the purpose of education to be to nurture each individual, regardless of background, to become a faithful communist. For decades, testing remained in the closed realms of clinical psychology only to regain a legitimate position after the demolition of the Soviet system in 1991. Since then, Russia has joined the global culture of standardised testing in spite of, or perhaps precisely due to, this historical context. For instance, Russian assessment experts central for the development of assessment policies and practices have gladly joined the ranks of international assessment experts of the OECD and the World Bank, and learned from them in order to reconnect with the worldwide scientific community to realign national testing with international policies and practices (see Piattoeva & Gurova 2018).

Standardised assessment in the context of international collaboration

The overall developments in standardised testing relied on and provided opportunities for international co-operation. When large-scale testing started to gain prominence in the 1960s, it constituted efforts to explore the feasibility of comparing educational achievements across countries and cultures. Additionally, these early studies experimented with the idea of examining the outcomes of education systems "to understand and improve education" (Hastedt 2020: 21). The first large-scale international studies were conducted in mathematics, as this subject started to enjoy growing political interest and was also considered to be the least culturally and linguistically sensitive. In the second wave of international assessments before the 1980s, the studies were nationally co-ordinated. This changed in the 1980s as responsibility for co-ordinating and leading studies on educational achievement was transferred to international collaborative organisations, which made decisions on sampling, scaling and instrument development. The IEA (International Association for the Evaluation of Educational Advancement) has conducted quantitative comparative assessments since 1958. The OECD's (Organisation for Economic Co-operation and Development) social indicator programmes were tried out during the 1970s and the International Indicators of Education Systems (INES) programme from 1988. Consequently, the studies became more standardised across the participating countries, however, with some possibilities for national adaptations (Hastedt 2020: 22–23.)

As we have illustrated in this section, assessment and evaluation have been fundamental to the functioning of public mass education across time and space. They manifest how evaluation and assessment are ingrained elements of the contemplation and functioning of education and its governance (see also Knudsen 2016). The historical threads of school inspection and standardised testing establish a ground for further expansion of assessment and evaluation practices locally, nationally and internationally. In order to understand this expansion, it is important to ask questions such as who can determine what is to be evaluated and how, how is data used and to whom it is made available, and in what format. Answers to these questions all give rise to the idea of the highly political nature of evaluation and evaluation data. This aspect may easily be concealed or forgotten by current discourses and practices that approach assessment and evaluation based on massive data collection as a technical matter, that is, as being neutral, objective and value-free information gathering.

Contemporary developments in assessment and societies at large

Many researchers have started to conceptualise societal and governance change in terms of evaluation due to its expansion and growing role. Accordingly, Dahler-Larsen (2012a) introduced the concept of "evaluation society", Neave (1998) the "evaluative state", and Power (1999; 2003) the "audit

society". In the sphere of education, concepts such as the "global testing culture" (Smith 2016) or the "global education reform movement" (Sahlberg 2016) have highlighted the increasing role of standardisation including that of assessment through testing. To explain how contemporary societies and education systems have come to be defined through evaluation, we must look at the political move from a centralised and top-down governed regulation system towards a decentralised and deregulated system that took place around the 1980s and 1990s as part of the changing governance paradigm and market-based education reforms. In addition, issues such as the rising interest and role of international organisations and transnational networks in national and local governance of education, as well as the recurring problematique of governance failure and a need for substitute tools of direct intervention and regulation, deserve further attention (Le Gales 2016).

Assessment and evaluation as part of changing decentralised governance

There is an increased demand for data about the workings of the public sector including education (Lawn 2013). The development is supported by transnational agencies (e.g. OECD; World Bank), which promote 'post-bureaucratic' models of governance as alternatives to the earlier 'bureaucratic–professional model' (cf. Maroy 2009). Globally, we are seeing a variety of hybrid policies and governance patterns that combine elements of both old and new models in creative ways, attesting to path-dependent trajectories. Whatever the case may be, it is important to understand how 'post-bureaucratic' models operate through an elusive frame that splits up and assigns tasks to an array of new instruments, and new intermediary bodies and actors, whose functions are legitimised through extensive legal modifications (see Neave 1998). This 'multiplication of the levels of oversight' (Neave 1998) explains the rise of new actors and hybrid government arrangements – for instance, specific evaluation agencies and networks that combine private and public actors and groups of scientific experts tasked with producing evaluation (cf. Kauko, Jaakko, Suominen, Centeno, Piattoeva & Takala 2018b).

The extent of evaluation and the associated capacity development facilitated by the international organisations (e.g. Kauko, Jaakko, Suominen, Centeno, Piattoeva & Takala 2018b), and the post-bureaucratic development of state governance, have contributed to a close connection between and even an interpenetration of international and national scales. This relates to the ways in which international framing, national inputs and the interaction between them form educational agendas (Centeno 2017). For instance, in Finland, Norway and Iceland, it was noted that while the actual documentary references to the OECD were not that important, the built-in networks, largely due to

the post-bureaucratic design of governance, afforded international organisations a deeper level of influence. This would entail definitions of the type of knowledge considered valuable to collect and use for decision-making (Ydesen, Kauko & Magnúsdottir forthcoming).

Decentralisation and deregulation diminished the state's regulative power and increased autonomy at local level and in schools. However, the steering power of the state did not disappear; the technique of exercising power changed. (e.g. Pitkänen 2019) Since then it has combined direct and indirect means of regulating. The exercise of indirect power or governing at a distance takes place through evaluative techniques such as target-setting, performance measurement, quality indicators and large-scale standardised testing. To illustrate, in the Nordic countries, the reform of decentralisation and deregulation during the 1980s and 1990s gave municipalities more autonomy (especially in Finland, Sweden and Iceland). But simultaneously, central government intensified evaluation of performance with inspections and standardized tests, signalling a form of recentralisation (Dovemark, Kosunen, Kauko, Magnúsdóttir, Hansen & Rasmussen 2018). However, in the seemingly homogeneous Nordic context, the Finnish education system, with school inspections discontinued and standardised tests remaining scarce despite similar decentralising reforms, stands out as an exception when compared to many other education systems in the region and beyond (Simola, Rinne, Varjo, Pitkänen & Kauko 2009; Kauko, Varjo & Pitkänen 2020; Dovemark et al 2018). A counterexample to this is England, for instance, where, despite efforts to emphasise school autonomy with the academy and free school reforms, school inspections have become even more pervasive, as they have been coupled with test-based accountability policies (see e.g. Kauko & Salokangas 2015; Wallenius 2020). The illustrative cases of different outcomes in reforms claiming to strengthen school or local-level autonomy shows how the production and use of evaluation data is intensifying in the wider context of convergent education policy reforms, which we explore further in the following.

Mapping convergent trends globally and nationally

Evaluation and assessment have become objects of policy and popular policy instruments, whereas the data acquired through evaluations have become a central means of governing and steering education at a distance, nationally and globally. Joint efforts on the Millennium Development Goals and later Sustainable Development Goals exemplify the common understanding described in the Incheon Declaration (UNESCO 2016: 66) by UNESCO, the World Bank Group and other UN organisations: "research and assessment culture is necessary at the national and international levels." International organisations have been proactive in promoting and conducting assessment and

evaluation, particularly in the form of quantitative standardised tests. A massive and perhaps the best known comparative assessment effort by an international organisation is the OECD's PISA survey, conducted regularly for almost two decades. The OECD Director of Education and Skills, Andreas Schleicher (2019: 4), sees PISA as "the world's premier yardstick for comparing quality, equity and efficiency in learning outcomes across countries, and an influential force for education reform." The process of forming PISA indicators is a complex web of negotiations and agreements (Carvalho 2012). The collected data are multifaceted and rich, and so are the main results reported. However, the focus in public debates is usually on the ordered performances of the countries, regions and economies taking part in the evaluation. It is, however, important to look behind the facile scores and rankings for more informative data, as well as to understand rankings as policy instruments that govern at a distance (Espeland & Sauder 2007).

The view promoted by international organisations and shared by the authorities of many countries is that evaluation data in education are beneficial for screening and improving education quality, ensuring accountability and raising economic competitiveness (Kauko, et al. 2018b; 2018c; Carvalho et al. 2020; Dovemark et al. 2019; Verger et al. 2019 Kamens & McNeely 2010). With respect to participation in international large-scale assessments, for instance, countries are motivated by a range of rationales, such as searching for evidence for policy, technical capacity-building, funding and aid, international relations, national politics, economic prerogatives and curriculum and pedagogy development (Addey & Sellar 2018). Moreover, it is both data collection and the act of participation that drive countries' interest in large-scale assessments such as PISA. The diverse and overlapping rationales for participation cannot be neatly mapped according to the participant country's level of income, manifesting the varying meanings and purposes that assessments acquire across contexts (Addey & Sellar 2018). However, the expansion of educational evaluation to middle- and low-income countries has been explicitly on the agenda of UNESCO, the World Bank and the OECD through discourses of "learning crisis" and "quality education" (e.g. Sriprakash, Tikly & Walker 2020). The expansion of evaluation has been undertaken through leverage gained with the combined effort of the aforementioned international organisations (Auld, Rappleye & Morris 2019), who help to disseminate new instruments such as "PISA for development" to new contexts (Addey & Gorur 2020).

In addition to promoting evaluation as a crucial education policy area, international organisations use evaluation data to produce policy recommendations. They create a vision of what is normal or desirable on highly complex and at times conflictual and sensitive issues in the field of education, but also far beyond. For instance, the World Bank conducted a massive evaluation exercise followed by a report entitled *Great Teachers in Latin America and the Caribbean Region* (Bruns & Luque 2015).

In this evaluation, 15,675 classrooms in 3,015 schools were observed in the region using a method that records academic activities, classroom management and teacher off-task times through short observations that simplify and de-contextualise classroom activity through standardised measurement (Bruns & Luque 2015). For example, students are considered uninvolved if they are "staring out of the window, resting his/her head on the desk, or sleeping" (World Bank Group 2015: 18).

Besides illustrating how evaluation and assessment tools are scaled up to produce panoramic and comparative views on a breadth of countries and education areas, the World Bank's evaluation tool indicates how assessment and evaluation as a policy area and a policy instrument overlap. The World Bank recruits individual countries to develop evaluation policies and practices particularly through programmes of building capacity in large-scale standardised testing and observation. Simultaneously it promotes the idea that evaluation is necessary and, as in the case of the Latin American and Caribbean regions, absolutely imperative to address the poor learning results documented in the tests that the World Bank supports. The World Bank argues that their observation method is particularly useful for documenting "differences in teacher effectiveness" across countries, thus relating student assessment results directly to teachers' pedagogical 'input' (rather than to socio-economic factors, for example) (see Bruns & Luque 2015: 132–133). The policy recommendations foreground the low quality of teachers and how this could be improved with recruitment, training and (fiscal) incentives. Moreover, the report suggests that the power of the Latin American teachers' unions poses a significant political challenge and an obstacle to education reform (Bruns & Luque 2015: 3). Here, data drawn from student performance tests is seen as a powerful tool to form alliances against trade unions (Bruns & Luque 2015: 324; co-author Marco Fernandez credited). The World Bank study and the ensuing report are examples of both how the policy area is broadened beyond student evaluation to destabilise trade unions and how the policy instrument (data) serves as a powerful tool in this endeavour.

Identifying national variation

Drawing on a large database on national assessments, Verger, Parcerisa and Fontdevila (2019) classify them according to the stated goals such as certification, diagnostics or monitoring in primary and lower secondary education across countries during the period 1995–2014. The main trends documented in the study are that first, the number of countries conducting sample or census-based national assessments has rapidly tripled to around thirty in the early 2000s and then plateaued, while the number of assessments has continued to rise. The use of assessment for certification, such as

selection and streaming, has remained around the same during the period analysed and there has been a moderate increase in the use of assessment for diagnostic purposes (without any formal implications). A striking increase has taken place in assessments for monitoring purposes, that is, using national large-scale assessments for measuring performance and progress. This category can explain most of the increase in assessment during the time period of interest. The monitoring usually focuses on schools and education systems or solely on education systems (Verger et al. 2019). Another comparative project, drawing on 200 interviews as well as documentary and observation data in Brazil, China and Russia found that evaluation and thus the ensuing data collection have become a self-reinforcing goal: there was a shared understanding at different administrative levels that collection of evaluation data would be key in solving education problems. This led to more data collection procedures forming an incessant cycle (Kauko et al 2018c; Piattoeva, Centeno, Suominen & Rinne 2018).

As we observe some general, convergent trends regarding assessment as objects and means of policy in and of education, the recent research shows that their actual implementation is conditional on local histories and discursive frames, preceding policy configurations, expertise, administrative regimes and the overall complexity and contingency of education policy and wider societal contexts (e.g. Kauko et al. 2018a; 2018c; Piattoeva & Gurova 2018). For instance, in some contexts, the introduction of national large-scale assessments became a component of the policies of test-based accountability (TBA). These policies deploy assessments to monitor teachers' performance and promote competitive pressures among schools. Moreover, they may have material, reputational, individual or collective consequences for teachers, administrators or local authorities according to the level of performance measured by externally administered standardised assessments that follow centrally-defined learning objectives. (Verger, Parcerisa & Fontdevila 2019.) In some countries adhering to TBA, national assessments are further complemented by inspections of adherence to nationally mandated education legislation (e.g. Russia, see Gurova 2019), thus showing further variations in the constitutive elements of TBA policies. At the same time, TBA policies *per se* are not the sole reason for the rise and proliferation of national assessments, as we show below. Moreover, test-based accountability systems operating on the bases of large-scale assessment data oscillate between high, middle and low stakes (cf. Maroy 2015).

As has been recently summarised by Diaz Rios (2020), "policy legacies shape (1) the extent to which education stakeholders regard tests as proper tools of accountability, (2) the incentives for actors to transfer and/or avoid the blame and consequences of poor education results, and (3) the power distribution among actors supporting or opposing TBA policies." Different configurations of these

legacies shape diametrically opposite policy outcomes, thus calling for contextually sensitive and relational analyses of assessment and evaluation policies. Clearly, implementing NLSA does not necessarily lead every country to develop test-based accountability (Diaz Rios 2020).

Policies are also shaped by discursive conditions. The rise of assessment and evaluation as major policy concerns are embedded in and promoted by the overall prerogative of "raising education quality," as we argued above, with quality becoming synonymous with and manifested by quantitative and nationally and internationally comparable assessment data. However, connotations of quality are not universal: Brazil, for instance, was found to exhibit a predisposition towards democratic ideals of "social quality" emphasising completion rates, universal access and retention rather than mere competition and accountability by results (Minina, Piattoeva, Centeno, Zhou & Candido 2018). Consequently, despite the adoption and general use of NLSAs, Brazil retained both a critical discursive space and an ambiguous policy with regard to accountability testing (ibid.; also Candido 2020).

Arising from these notions, a seemingly uniform national large-scale assessment data set can be used for many different purposes, such as improving international reputation (Brazil) or increasing state control (Russia and China) (Kauko et al. 2018b). Evaluation data can be used in political struggles between government and opposition (Santos & Kauko 2020). A typical way in the Nordic countries in the context of decentralisation has been to reassert central power (Dovemark et al. 2018), as shown above. This demonstrates how the power vested in evaluation data makes it attractive for use for purposes other than directly supporting education-related aims (Piattoeva 2015). In Brazil, for instance, the third sector movements, such as *Todos pela Educação*, were able to strengthen their position and initiate policy change with the help of openly available evaluation data. Chinese education experts could find more room for action due to the increased importance of data. The Russian government was able to tighten their grip on local actors with the help of evaluation data. (Kauko et al. 2018b.) The Finnish Education Evaluation Council was formed as a merger of former evaluating agencies, and has since been given a more prominent role in the strategic management of Finnish education evaluation (Kauko et al. 2020). Evaluation data creates space for new players in education policy and beyond, and simultaneously enhances the role of actors responsible for the production and analysis of data. These two phenomena work hand in hand. The first relates to the increased importance of evaluation data and how it is used as a means of governance. The second relates to how actors wielding this power become empowered in the process.

On the political nature of evaluation data and its use

Alongside the apparently growing emphasis on the collection and use of evaluation data, its actual availability and deployment are open to diverse interpretations and a range of practices of use and non-use (e.g. Waldow 2010; Santos & Kauko 2020; Piattoeva et al. 2018). This can be illustrated by cases of heated and controversial political debates on the public availability of school-level performance evaluation data in diverse contexts, around the issues of how, for whom and for what purposes data should be made available (Karsten, Visscher & De Jong 2001; Wallenius 2020). In some contexts, e.g. in Sweden, public availability has been fostered by the principle of granting parents equal access to official information, whereas in other contexts, it has been presented as providing better quality through competition (West & Pennell 2000). This has been especially the case in the British and American educational cultures (e.g. the USA, UK, Canada, Australia), where school-specific performance tables (school ranking lists, league tables) are used to inform school choice. Elsewhere, by contrast, keeping evaluation data out of public view was supported to prevent the naming and shaming of schools and the risk of social segregation through public school rankings in the media (Wallenius 2020). Thus, national policies on data availability depend on the country's deep-seated institutionalised perceptions on marketisation, competition and adherence to test-based accountability in educational policymaking.

In Europe, the policy of publishing comparable data on individual schools was first introduced in the UK in the early 1990s (West & Pennell 2000). Since the 2000s, policies supporting data availability have been spreading in other contexts such as Sweden, Denmark, Norway, the Netherlands and Estonia (Eurydice 2009; Wallenius 2020). Several reasons for this development can be noted. Recent technological advances in digitalisation and datafication have enabled the data collection, management, analysis and visualization in significantly more convenient, modifiable and user-amenable ways. The rise of computer-based assessments and the availability of statistical software have furthered the use of school evaluation indicators. In particular, the use of value-added modelling, which aims to provide information on schools' and pupils' relative performance by controlling for various background variables (e.g. pupils' prior achievement or parents' socio-economic background) as covariates, increased significantly in the 2000s (Levy, Brunner, Keller & Fischbach 2019). Additionally, there is a rising trend of data publication and visualization through the official internet web-portals of national governments, enabling public comparisons of schools (Wallenius 2020). This also relates to the ideological change that has driven data publicity. Here, we point out the role of accountability and transparency as policies and political narratives through which data availability has been intensified since the waves of negative "PISA shocks" in Sweden, Norway and Denmark in

the early 2000s (Egelund 2008; Ringarp 2016; Tveit 2014). As Dubnick (2014) argues, the power in accountability lies in the political narratives that promise a combination of more efficient control, better performance and improved justice and democracy. Combined with the longstanding ideal of governance transparency characteristic of the Nordic context (see Erkkilä 2012), it is easy to see how many policymakers have shown a keen interest in data availability and school comparisons, despite the critique and warnings of its unwanted consequences.

To end this section, we turn to two cases in the Nordic region to highlight once again the relevance of socio-historical context. First, Finland makes an interesting case of divergence amidst the emergence of a global testing culture (Smith 2016). Even if governance transparency is appreciated as highly in Finland as throughout the Nordic countries, the Finnish school evaluation culture differs significantly from that of the other Nordic countries. At the level of comprehensive education, school rankings are widely discouraged and the school evaluation data mainly serve educational officials, and not parents or the wider public (Simola et al. 2009). Safeguards against the evaluation data being transformed into national school rankings include a preference for sample-based testing (Wallenius 2020). For the last 20 years, Finland's success in PISA has undoubtedly reduced the pressure to revise the policy guidelines on school evaluation. Nevertheless, it is more appropriate to understand how these guidelines are institutionalised socio-historically both in formal school evaluation practices and in policy discourses that draw on school autonomy and teachers' professionalism as opposed to external accountability (Wallenius, Juvonen, Hansen & Varjo 2018; Wallenius 2020).

Second, an episode from December 2019 in Sweden illustrates the political nature of school evaluation in general and of data availability in particular. Sweden has been a front runner among the Nordic countries in data availability, and has published the evaluation results on a school-by-school basis since 2001. Unexpectedly the Administrative Court of Gothenburg ruled several school evaluation statistics (e.g. pupils' grading, throughput, composition) at independent schools in Sweden to be classified information under business privacy law (SNAE 2020a). The new legal interpretation was not confirmed by the school evaluation agency responsible, the Swedish National Agency for Education. They soon announced the removal of the data from the web portals and also from public schools in the name of equality. This decision caused a heated debate in Sweden and culminated in August 2020 in an open letter signed by several Swedish academics, in which the signatories appealed for the data to be reinstated for use in Swedish schools and their development (Svenska Dagbladet 2020). Finally, in October 2020, a new court ruling ordered that the school statistics be once again made publicly available, while leaving the future publication policy undecided (SNAE 2020b). The

Swedish case shows how even long-established practices in evaluation may suddenly become politicised and be challenged and re-evaluated.

Conclusion

The chapter deployed post-structural theorising drawing on the critical traditions of sociology and politics of education to examine assessment and evaluation as objects and means of policy. It illustrated, on the one hand, a general cross-national convergence in the purposes and means of assessment and evaluation in formal mass education, and on the other hand examined the diverse ways in which countries have incorporated these into their national policies. Assessment and evaluation constitute a distinct area of education policy that has attracted increasing attention among international and national actors across countries; it is also a policy instrument that enables these actors to act on the evaluated entity directly and at a distance. The contextuality of assessment policies and the application of assessment as policy tools arises from the perpetuation of historically embedded assessment cultures (cf. Alarcon & Lawn 2018) that can be disentangled by examining concrete practices of data collection and use for evaluation.

As a policy area, assessment and evaluation are increasingly seen as crucial for educating students who are "better fitted" to meet the demands of the global economy. Moreover, they are seen as imperative for monitoring and thus improving the quality of education, where this is commonly defined as meeting external demands on education rather than those intrinsic to the diverse interests, values and aspirations of educated and educators. These external demands are defined, for instance, by narratives of the "knowledge economy" that identify national education as a vehicle of economic competitiveness relying on a skilled workforce. Yet assessment and evaluation serve and are simultaneously an easy fit with a range of political programmes and post-bureaucratic discourses, including different forms of accountability, transparency, equality, efficient control and timely intervention. As we have illustrated, these political agendas and historically shaped cultures play a prominent role in how assessment and evaluation are understood and put to work globally and across specific national contexts. Yet countries are not uniform in their political agendas and adherence to one or a combination of the aforementioned political agendas. Nevertheless, they increasingly trust and rely on assessment and evaluation, spurring on the further production and use of data, and often making it publicly available.

Conceptualising assessment and evaluation as a policy instrument helped us to further explore their political nature, focusing on how they enable a governing actor to shape the conduct of the entity

evaluated at a distance. This can be done, for instance, by setting achievement targets and offering material or symbolic rewards or sanctions for various levels of achievements. As a policy instrument, assessment and evaluation also exercise power through normalization on both individual and collective levels, that is, by defining the limits of educational normality or deviance, determining which examples should be emulated, and setting a trajectory to follow − for instance, through rankings common to international standardised assessments. Moreover, the very character of assessment and evaluation as standardised, quantitative and increasingly large-scale confers on them the status of objective, authoritative and comprehensive 'knowledge'. Finally, assessment and evaluation practices that produce data oscillate between notions of "numbers speaking for themselves" and numbers as inviting and being open to further interpretation. The latter enables assessment data to be used for diverse political agendas, both existing and emerging.

The production and use of assessment data carry risks and raise further questions. As a policy area and instrument that attempts to cope with the complexity of educational phenomena, it ends up creating a gap between the representation of the phenomena by the instrument deployed and the complexity of the reality that it strives to capture and intervene in (Rottenburg & Merry 2015). This gap has been a constant point of critique by various researchers and actors in the field (cf. Piattoeva 2021). The main criticism of the PISA rankings, for instance, has been their lack of attention to the sociological and historical factors contributing to the reported cross-national variation (e.g. Simola 2005; Hwa 2019). The World Bank's watchword economic growth, and the strong statistical claims that improvements in global learning assessments such as PISA will lead to higher GDP growth, have equally been debunked (Komatsu & Rappleye 2017). These debates are an important reminder that education evaluation and the data it produces find relevance and use in other policy sectors beyond education. In addition, the goals set for education that spur on the desire for more assessment and evaluation, and the data they produce, are shaped by agendas that are not specifically of or for education. In sum, the data and political agendas underpinning the evolution of assessment and evaluation are evolving and are subject to critique along with changing political conditions and prerogatives. Historical sensitivity, comparative approaches and examining the phenomenon from less familiar geographical vantage points − as we have done in the chapter − remind us that assessment and evaluation are far from uncontested, fixed or uniform.

Since the 2000s there has been a growing tendency to approach evaluation practices and data in a more holistic way beyond cognitive learning results or general statistics on school resources. These developments have attempted to address some of the critiques mentioned above, e.g. foregrounding the social and emotional aspects of learning in addition to cognitive results or going beyond simplistic

school rankings. However, a more fundamental question that goes beyond the focus on more varied, more accessible, more elaborate, more frequent and more usable assessment and evaluation (data), remains to be answered. This question is: does the proliferation of assessment policies and practices truly enable and embrace everyone's opportunities to live a 'good life', rather than support actors striving for power, profit or prestige?

References

Addey, Camilla, and Radhika Gorur (2020), "Translating PISA, Translating the World," *Comparative Education*, 56 (4): 547–64. DOI: 10.1080/03050068.2020.1771873

Addey, Camilla, and Sam Sellar (2018), "Why Do Countries Participate in Pisa? Understanding the Role of International Large-Scale Assessments in Global Education Policy," in Antoni Verger, Mario Novelli and Hülya K. Altinyelken (eds), *Global Education Policy and International Development*, 2nd ed., 97–117, London: Bloomsbury.

Alarcón Lopez Cristina, and Martin Lawn eds. (2018), *Assessment Cultures: Historical Perspectives*. Studia Educationes Historica. Vol. 3. Bern: Peter Lang.

Auld, Euan, Jeremy Rappleye, and Paul Morris (2019), "PISA for Development: How the OECD and World Bank Shaped Education Governance post-2015," *Comparative Education*, 55 (2): 197–219. DOI: 10.1080/03050068.2018.1538635

Ball, Stephen J. (2003), "The Teacher's Soul and the Terrors of Performativity," *Journal of Education Policy, 18* (2): 215–228. DOI: 10.1080/0268093022000043065

Berry, Rita, and Bob Adamson (2011), "Assessment Reform Past, Present and Future" in Rita Berry and Bob Adamson (eds), *Assessment Reform in Education Policy and Practice*, 3–14, Dordrecht: Springer Netherlands.

Bruns, Barbara, and Javier Luque (2015), *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean.* Washington, DC: World Bank. https://openknowledge.worldbank.org/handle/10986/20488

Candido, Helena Hinke Dobrochinski (2020), "Datafication in Schools: Enactments of Quality Assurance and Evaluation Policies in Brazil," *International Studies in Sociology of Education*, 29 (1–2): 126–57. DOI: 10.1080/09620214.2019.1656101

Carvalho, Luís Miguel (2012), "The Fabrications and Travels of a Knowledge-policy Instrument," *European Educational Research Journal*, 11 (2): 172–88.

Carvalho, Luís Miguel, Estela Costa, and Carlos Sant'Ovaia (2020), "Depicting the Faces of Results-Oriented Regulatory Processes in Portugal: National Testing in Policy Texts," *European Educational Research Journal*,19 (2): 125–41. https://doi.org/10.1177/1474904119858799Centeno, Vera G. (2017), *The OECD's Educational Agendas – Framed from Above, Fed from Below, Determined in Interaction. A Study on the Recurrent Education Agenda*, Berlin: Peter Lang.

Centeno, Vera. G., Jaakko Kauko, and Helena Hinke Dobrochinski Candido (2017), "Quality Assurance and Evaluation through Brazilian lenses: an exploration into the validity of umbrella concepts," *Comparative Education*, 54 (2): 132–58. DOI: 10.1080/03050068.2017.1348084

Dahler-Larsen, Peter (2012a), *The Evaluation Society*, Stanford, CA: Stanford University Press.

Dahler-Larsen, Peter (2012b), "Constitutive effects as a social accomplishment: A qualitative study of the political in testing," *Educational Inquiry*, 3 (2): 171–86.

Dahler-Larsen, Peter (2019), *Quality. From Plato to Performance*. Cham: Springer Nature / Palgrave Macmillan.

Diaz Rios, Claudia Milena (2020), "The Role of Policy Legacies in the Alternative Trajectories of Test-Based Accountability," *Comparative Education Review*, 64 (4): 619–641.

Dovemark, Marianne, Sonja Kosunen, Jaakko Kauko, Berglind Magnúsdóttir, Petteri Hansen, and Palle Rasmussen (2018), "Deregulation, Privatisation and Marketisation of Nordic Comprehensive Education: Social Changes Reflected in Schooling," *Education Inquiry* 9 (1): 122–41.

Dubnick, Melvin J. (2014), "Accountability as a Cultural Keyword," in Mark Bovens, Robert E. Goodin, and Thomas Schillemans (eds), *The Oxford Handbook of Public Accountability*, 23–28, Oxford: Oxford University Press. DOI: 10.1093/oxfordhb/9780199641253.001.0001

Egelund, Niels (2008), "The Value of International Comparative Studies of Achievement – A Danish Perspective," *Assessment in Education: Principles, Policy & Practice*, 15 (3): 245–51. DOI: 10.1080/09695940802417400

Erkkilä, Tero (2012), *Government Transparency: Impacts and Unintended Consequences*. Basingstoke: Palgrave.

Eurydice (2009), *National testing of pupils in Europe: Objectives, organization and use of results.* EACEA: Eurydice, Brussels: European Commission DOI: 10.2797/18294.

Evertsson, Jakob (2015), "History, Nation and School Inspection: The Introduction of Citizenship Education in Elementary Schools in Late Nineteenth-Century Sweden," *History of education (Tavistock)* 44 (3): 259–273.

Foucault, Michel (1982), "The Subject and Power," *Critical Inquiry*, 8 (4): 777–795.

Foucault, Michel (2000), "Governmentality," in James D. Faubion (ed), *Power. Essential Works of Foucault 1954-1984, 2nd ed*, 201–222, New York: The New Press.

Grek, Sotiria, Christian Maroy, and Antoni Verger (2020), "Introduction," in Sotiria Grek, Christian Maroy, and Antoni Verger (eds), *World Yearbook of Education 2021: Accountability and Datafication in the Governance of Education,* 1-22*, Milton: Taylor and Francis.*

Grek Sotiria, and Jenny Ozga (2010), "Re-Inventing Public Education: The New Role of Knowledge in Education Policy Making," *Public Policy and Administration*, *25* (3): 271–288. DOI: 10.1177/0952076709356870

Grimaldi, Emiliano (2020), *An Archaeology of Educational Evaluation: Epistemological Spaces and Political Paradoxes.* Milton: Routledge.

Gunter, Helen M., David Hall, and Colin Mills (2015), "Consultants, Consultancy and Consultocracy in Education Policymaking in England," *Journal of Education Policy*, 30 (4): 518–539. DOI: 10.1080/02680939.2014.963163

Gurova, Galina (2019), *Quality Assurance and Evaluation as a Mode of Local Education Governance: The Case of Russian Schools*. Doctoral diss., Faculty of Education and Culture, Tampere University, Tampere: Tampere University Press.

Hacking, Ian (1991), "How Should We Do the History of Statistics?" in Graham Burchell, Colin Gordon, and Peter Miller (eds), *Foucault Effect: Studies in Governmentality*, 181–196, Chicago: University of Chicago Press.

Hastedt, Dirk (2020), "History and Current State of International Student Assessment," in Heidi Harju-Luukkainen, Nele McElvany, and Justine Stang (eds), *Monitoring Student Achievement in the 21st Century: European Policy Perspectives and Assessment Strategies*, 21–37, Cham: Springer International Publishing AG.

Harvey, L. (2004). “Analytic quality glossary. Quality Research International.” Available online: http://www.qualityresearchinternational.com/glossary/quality.htm (accessed June 14, 2021)

Hwa, Yue-Yi (2019), *“Teacher Accountability Policy and Sociocultural Context: A Cross-Country Study Focusing on Finland and Singapore,”* Doctoral diss., University of Cambridge. DOI: 10.17863/CAM.55349

James, David C. and Brian Davies (2009), “The Genesis of School Inspection in South East Wales 1839–1843: Issues of Social Control and Accountability”, *History of education (Tavistock)* 38 (5): 667–80.

Kamens, David H., and Connie L. McNeely (2010), “Globalization and the Growth of International Educational Testing and National Assessment,” *Comparative Education Review* 54 (1): 5–25.

Kauko, Jaakko, Risto Rinne, and Tuomas Takala (2018c), “Conclusion,” in Jaakko Kauko, Risto Rinne & Tuomas Takala (eds), *Politics of Quality in Education: A Comparative Study of Brazil, China, and Russia*, 180–90, London: Routledge.

Kauko, Jaakko, and Maija Salokangas (2015), “The Evaluation and Steering of English Academy Schools Through Inspection and Examinations: National Visions and Local Practices,” *British Educational Research Journal*, 41 (6): 1108–24.

Kauko, Jaakko, Olli Suominen, Vera G. Centeno, Nelli Piattoeva, and Tuomas Takala (2018b), “Established and Emerging Actors in The National Political Arenas,” in Jaakko Kauko, Risto Rinne and Tuomas Takala (eds), *Politics of Quality in Education: A Comparative Study of Brazil, China, and Russia*, 71–90, London: Routledge.

Kauko, Jaakko, Tuomas Takala, and Risto Rinne (2018a), “Comparing Politics of Quality in Education,” in Jaakko Kauko, Risto Rinne and Tuomas Takala (eds), *Politics of Quality in Education: A Comparative Study of Brazil, China, and Russia*, 1–17, London: Routledge.

Kauko, Jaakko, Janne Varjo, and Hannele Pitkänen (2020), “Quality and Evaluation in Finnish Schools,” *Oxford Research Encyclopedia of Education.* Oxford University Press.

Kellaghan, Thomas, Daniel L. Stufflebeam, and Lori A. Wingate (2003), “Introduction”, in Thomas Kellaghan and Daniel L. Stufflebeam (eds), *International Handbook of Educational Evaluation*, 1–6, Dordrecth: Kluwer Academic Publishers.

Kipnis, Andrew B. (2008), "Audit cultures: Neoliberal Governmentality, Socialist Legacy, or Technologies of Governing?" *American Ethnologist*, 35 (2): 275–289. DOI: 10.1111/j.1548-1425.2008.00034.x

Knudsen, Andrew T. (2016), "Profession, 'Performance', and Policy: Teachers, Examinations, and the State in England and Wales, 1846–1862," *Paedagogica historica*, 52 (5): 507–524.

Komatsu, Hikary and Jeremy Rappleye (2017), "A New Global Policy Regime Founded on Invalid Statistics? Hanushek, Woessmann, PISA, and Economic Growth," *Comparative Education*, 53 (2): 166–191. DOI: 10.1080/03050068.2017.1300008

Lascoumes, Pierre, and Patric Le Gales (2007), "Introduction: Understanding Public Policy through its Instruments – From the Nature of Instruments to the Sociology of Public Policy Instrumentation." *Governance*, 20 (1): 1–21.

Lawn, Martin (2013), *The Rise of Data in Education Systems. Collection, Visualization and Use*. Oxford: Symposium Books.

2013Le Galès, Patric (2016), "Performance measurement as a policy instrument," *Policy Studies*, 37 (6): 508–520.

Levy, Jessica, Martin Brunner, Ulrich Keller, and Antoine Fischbach (2019), "Methodological Issues in Value-Added Modeling: An International Review From 26 Countries," *Educational Assessment, Evaluation and Accountability*, 31 (3): 257–287. DOI: 10.1007/s11092-019-09303-w.

Maroy, Christian (2009), "Convergences and Hybridization of Educational Policies Around 'Post-Bureaucratic' Models of Regulation," *Compare: A Journal of Comparative and International Education*, 39 (1): 71–84. DOI: 10.1080/03057920801903472.

Maroy, Christian (2015), "Comparing Accountability Policy Tools and Rationales," in Hans-Georg Kotthoff and Eleftherios, Klerides (eds), *Governing Educational Spaces: Knowledge, Teaching, and Learning in Transition*, 35–56, CESE Rotterdam: Sense Publishers. DOI: 10.1007/978-94-6300-265-3_3

Minina, Elena, Nelli Piattoeva, Vera G. Centeno, Xingguo Zhou, and Helena Hinke Dobrochinski Candido (2018), "Transnational Policy Borrowing and National Interpretations of Educational Quality in Russia, China, and Brazil," in Maia Chankseliani, and Iveta Silova (eds), *Comparing post-*

socialist transformations: purposes, policies, and practices in education, 27–44, Oxford Studies in Comparative Education. Oxford, U.K.: Symposium Books.

Neave, Guy (1998), "The Evaluative State Reconsidered," *European Journal of Education,* 33 (3): 265–284.

Ozga, Jenny, Peter Dahler-Larsen, Christina Segerholm, and Hannu Simola, eds. (2011), *Fabricating Quality in Education. Data and governance in Europe*, London: Routledge.

Olssen, Mark, and Michael A. Peters (2005), "Neoliberalism, Higher Education and the Knowledge Economy: From The Free Market to Knowledge Capitalism," *Journal of Education Policy*, 20 (3): 313–345. DOI: 10.1080/02680930500108718

Piattoeva, Nelli (2021), "Numbers and Their Contexts: How Quantified Actors Narrate Numbers and Decontextualization," *Educational Assessment, Evaluation and Accountability*, online first.

Piattoeva, Nelli (2015), "Elastic Numbers: National Examinations Data as a Technology of Government," *Journal of Education Policy,* 30 (3): 316–334.

Piattoeva, Nelli, and Galina Gurova (2019), "Domesticating International Assessments in Russia: Historical Grievances, National Values, Scientific Rationality and Education Modernization," in Lopez Christina Alarcón, and Martin Lawn (eds), *Assessment Cultures: Historical Perspectives*, 87–110, Frankfurt am Main: Peter Lang.

Piattoeva, Nelli, Vera G. Centeno, Olli Suominen, and Risto Rinne (2018), "Governance by Data Circulation? The Production, Availability, and Use of National Large-Scale Assessments Data," in Kauko, Jaakko, Risto Rinne, and Tuomas Takala (eds), *Politics of Quality in Education: A Comparative Study on Brazil, China, and Russia*, 115–136, London: Routledge.

Pitkänen, Hannele (2019), "Arviointi, tieto ja hallinta. Peruskoulun paikallisen arvioinnin genealogia," [Evaluation, Knowledge and Power: Genealogy of Local Quality Evaluation in the Field of Comprehensive Education], Doctoral diss., Faculty of Educational Sciences, Helsinki Studies in Education 50. Helsinki, Finland: University of Helsinki.

Power, Michael (1999), *The audit society: Rituals of Verification*. Oxford: Oxford University Press.

Power, Michael (2003), "Evaluating the Audit Explosion," *Law & Policy*, 25 (3): 185–202.

Ringarp, Johanna (2016), “PISA Lends Legitimacy: A Study of Education Policy Changes in Germany and Sweden after 2000,” *European Educational Research Journal*, 15 (4): 447–461. DOI: 10.1177/1474904116630754

Rose, Nikolas, and Peter Miller (2010), “Political Power Beyond the State: Problematics of Government,” *The British Journal of Sociology,* 61 (1): 271–303.

Rose, Nikolas, Pat O’Malley, and Mariana Valverde (2006), “Governmentality,” *Annual Review of Law and Social Science* 2: 83–104. DOI: 10.1146/annurev.lawsocsci.2.081805.105900

Rottenburg, Richard, and Sally Engle Merry (2015), “A World of Indicators: The Making of Governmental Knowledge Through Quantification,” in Richard Rottenburg, Sally Engle Merry, Sung-joon Park, and Johanna Mugler (eds), *The World of Indicators: the Making of Governmental Knowledge through Quantification*, 1–33, Cambridge: Cambridge University Press.

Saari, Antti (2011), ”Kasvatustieteen tiedontahto. Kriittisen historian näkökulmia suomalaiseen kasvatuksen tutkimukseen,” (Education and Will to Knowledge. A Critical History of Educational Research in Finland), Research in Educational Sciences 55, Jyväskylä: Finnish Educational Research Association.

Saari, Antti (2013), “Tilastollinen järkeily ja oppilasarviointi suomalaisen kasvatustieteen historiassa Ian Hackingin tieteenfilosofian näkökulmasta,” [Statistical Reasoning and Pupil Assessment in Finnish History of Educational Sciences from the Perspective of Ian Hackings Philosophy of Sciences], Kasvatus & Aika, 5–23.

Sahlberg, Pasi (2016), “The Global Educational Reform Movement and Its Impact on Schooling,” in Karen Mundy, Andy Green, Bob Lingard, and Antoni Verger (eds), *The Handbook of Global Education Policy*, 128–144, West Sussex: Wiley-Blackwell.

Santos, Íris, and Jaakko Kauko (2020), “Externalisations in the Portuguese Parliament: Analysing Power Struggles and (de-)Legitimation with Multiple Streams Approach,” *Journal of Education Policy*, DOI: 10.1080/02680939.2020.1784465

Schleicher, Andreas (2019), *PISA 2018: Insights and Interpretations*. Paris: OECD.

Simola, Hannu (2005), “The Finnish Miracle of PISA: Historical and Sociological Remarks on Teaching and Teacher education,” *Comparative Education*, 41 (4): 455–470.

Simola, Hannu, Risto Rinne, Janne Varjo, Hannele Pitkänen, and Jaakko Kauko (2009), ”Quality Assurance and Evaluation (QAE) in Finnish Compulsory Schooling: A National Model or Just Unintended Effects of Radical Decentralisation?” *Journal of Education Policy*, 24 (2): 163–178.

Sjoerd, Karsten, Andrie Visscher, and Tim de Jong (2001), “Another Side to the Coin: The Unintended Effects of the Publication of School Performance Data in England and France,” *Comparative Education*, 37 (2): 231–242.

Smith, William C. eds. (2016), *The Global Testing Culture: shaping education policy, perceptions, and practice*, Oxford: Symposium books.

SNAE (2020a), “Changed privacy policy affects access to certain statistics,” *Press release by the Swedish National Agency for Education*, June 26. 2020. Available Online: https://www.skolverket.se/skolutveckling/statistik/arkiverade-statistiknyheter/statistik/2020-06-26-forandrad-sekretesspolicy-paverkar-tillgang-till-viss-statistik (accessed June 18, 2021).

SNAE (2020b), “Previous school statistics will be available again,” *Press release by the Swedish National Agency for Education*, October 8. Available Online: https://via.tt.se/pressmeddelande/tidigare-skolstatistik-blir-tillganglig-igen?publisherId=743270&releaseId=3284059&fbclid=IwAR1pAjf5sAsbHgJFLSkaW5yc69YTtVyK-Ht-4giM0i-hvuP2tGQPgFpz2U0 (accessed June 18, 2021).

Sriprakash, Arathi, Leon Tikly, and Sharon Walker (2020), “The Erasures of Racism in Education and International Development: Re-Reading the ‘Global Learning Crisis,” *Compare: A Journal of Comparative and International Education*, 50 (5), 676–692. DOI: 10.1080/03057925.2018.1559040

Suominen, Olli, Vera G. Centeno, Galina Gurova, Johanna Kallo, and Xingguou Zhou (2018), “Historical Paths to Shared Interest in Quality Assurance and Evaluation,” in Jaakko Kauko, Risto Rinne, and Tuomas Takala (eds), *Politics of Quality in Education: A Comparative Study of Brazil, China, and Russia*, 44–70. London: Routledge.

Svenska Dagbladet. (2020). “En hemlig skola röjer det orimliga,” [A secret school reveals the unreasonable], August 30. Available Online: https://www.svd.se/en-hemlig-skola-rojer-det-orimliga (accessed June 18, 2021)

Tveit, Sverre (2014), “Educational Assessment in Norway,” *Assessment in Education: Principles, Policy & Practice*, 21 (2): 221–237. DOI: 10.1080/0969594X.2013.830079

Unesco (2016), "Education 2030: Incheon Declaration and Framework for Action for the implementation of Sustainable Development Goal 4: Ensure inclusive and equitable quality education and promote lifelong learning," Available Online: https://unesdoc.unesco.org/ark:/48223/pf0000245656 (accessed February 11, 2021).

Varjo, Janne, Hannu Simola, and Risto Rinne (2016), *Arvioida ja hallita—peräänkatsomisesta informaatio-ohjaukseen suomalaisessa koulupolitiikassa*, [To evaluate and govern—From "looking after" to management by data in Finnish education politics]. Research in Educational Sciences 70. Jyväskylä: Finnish Educational Research Association.

Vedung, Evert (2010), "Four waves of evaluation diffusion," *Evaluation*, 16 (3), 263–277.

Verger, Antoni, Lluís Parcerisa, and Clara Fontdevila (2019), "The Growth and Spread of Large-Scale Assessments and Test-Based Accountabilities: A Political Sociology of Global Education Reforms," Educational Review (Birmingham) 71 (1): 5–30. DOI: 10.1080/00131911.2019.1522045

Waldow, Florian (2010), "Der Traum vom "Skandinavisch Schlau werden". Drei Thesen zur Rolle Finnlands als Projektionsfläche in der Gegenwärtigen Bildungsdebatte," Zeitschrift für Pädagogik, 56 (4): 497–511.

Wallenius, Tommi, Sara Juvonen, Petteri Hansen, and Janne Varjo (2018), "Schools, Accountability and Transparency – Approaching the Nordic School Evaluation Practices Through Discursive Institutionalism," *Nordic Journal of Studies in Educational Policy*, 4 (3), 133–143. DOI: 10.1080/20020317.2018.1537432

Wallenius, Tommi (2016), Oppimistulosten kansallisen arvioinnin historiallinen institutionaalistuminen Suomessa ja Ruotsissa," [National Testing of Pupils in Finland and Sweden in Light of Historical Institutionalisation] in Heikki Silvennoinen, Mira Kalalahti and Janne Varjo (eds), *Koulutuksen tasa-arvon muuttuvat merkitykset. Kasvatussosiologian vuosikirja 1*, 99–131, Research in Educational Sciences 73. Helsinki: The Finnish Educational Research Association.

Wallenius, Tommi (2020), *Schools, Performance and Publicity: Contrasting the Policy on Publicising School Performance Indicators in Finland with the other Nordic Countries*. Doctoral diss. Faculty of Educational Sciences, Helsinki Studies in Education, no. 92. Helsinki: University of Helsinki.

West, Anne, and Hazel Pennell (2000). "Publishing School Examination Results in England: Incentives and Consequences," *Educational Studies*, 26 (4), 423–436.

World Bank Group (2015), “Conducting classroom observations,” Available Online:https://documents.worldbank.org/curated/en/790221467997639302/pdf/97904-WP-Box391498B-PUBLIC-WB-Stallings-web.pdf (accessed February 2, 2021).

Ydesen, Christian, Jaakko, Kauko, and Magnúsdóttir, Berglind Rós (forthcoming). “The OECD and the Field of Knowledge Brokers in Danish, Finnish, and Icelandic Education Policy,” in Kirsten Sivesind, Berit Karseth, and Gita Steiner-Khamsi (forthcoming) *Evidence and Expertise in Nordic Education Policies: A Comparative Network Analysis from the Nordic Region*. Palgrave.

Zhou, Xingguou (2019), *Changed and unchanged: the transformation of educational policies on assessment and evaluation in China*. Doctoral diss. Turku: University of Turku.

.