# Reforming governance through policy instruments: how and to what extent standards, tests and accountability in education spread worldwide

**Antoni Verger*, Clara Fontdevila, and Lluís Parcerisa**

All of *Department of Sociology, Universitat Autònoma de Barcelona, Spain*

In the last decades, most countries have adopted data-intensive policy instruments aimed at modernizing the governance of education systems, and strengthening their competitiveness. Instruments such as national large-scale assessments and test-based accountabilities have disseminated widely, to the point that are being enacted in countries with very different administrative traditions and levels of economic development. Nonetheless, comparative research on the trajectories that governance instruments follow in different institutional and socio-economic contexts is still scarce. On the basis of a systematic literature review (n=158), this paper enquires into the scope and modalities of educational governance change that national large-scale assessments and test-based accountability instruments have triggered in a broad range of institutional settings. The paper shows that, internationally, educational governance reforms advance through path-dependent and contingent processes of policy instrumentation that are markedly conditioned by prevailing politico-administrative regimes. The paper also reflects on the additive and evolving nature of educational governance reforms.

**Keywords:** accountability; decentralization; national assessments; historical institutionalism; standards; new public management; policy instruments

## Introduction

New forms of regulatory governance strongly rely on data-intensive policy instruments. These 'new' policy instruments are adopted on the top of more

* Corresponding author's email: E-mail: antoni.verger@uab.cat

traditional fiscal and legislative instruments in an attempt of steering increasingly fragmented and multi-layered policy systems more effectively (Wilkins & Olmedo, 2019). Data-intensive policy instruments are technically complex and sophisticated in design. Their main functions are to collect new forms of information on public sector performance, and to manage public services' conduct at a distance (Scott, 2000). In education, regulatory governance has meant the adoption of national large-scale assessments, test-based accountability, and explicit learning standards.

National large-scale assessments (NLSAs), which usually rely on the external evaluation of students' learning through standardized tests, are the governance instrument that has spread fastest in education systems recently. In the last two decades, the number of NLSAs being enacted globally has expanded exponentially, with their presence being especially important in OECD and middle-income countries (Ramirez, Schoffer, & Meyer, 2018; Verger, Parcerisa, & Fontdevila, 2018). Furthermore, NLSAs, beyond a data collection device, have become an intrinsic component of test-based accountability systems (TBA). The assemblage of NLSAs, standards and accountability constitutes 'a coherent and effective political dispositif' (cf. Ball, Junemann, & Santori 2017, p. 4) that, according to international data sources, is being increasingly enacted to monitor teachers' performance and promote competitive pressures among schools (Teltemann & Jude, 2018; Verger et al., 2018). In countries with TBA systems in place, school actors face consequences of a different nature (material, reputational, individual, collective, etc.) according to their levels of performance and adhesion to centrally-defined learning standards.

The origin of regulatory governance in education dates back to the 1980s, when mainly Anglo-Saxon countries introduced structural reforms in public administration following the tenets of neoliberalism. These reforms attempted to introduce marketization and privatization in the delivery of public services and to this purpose governments, among other measures, started publishing school rankings on the basis of NLSAs results. Later on, with the intensification of economic globalization, other countries started adopting similar policies as a way to monitor and strengthen the competitiveness of their educational systems. New governance instruments have often travelled as part of broader education reform packages that also promote decentralization, school autonomy and the diversification of school provision. Concepts such as the Global Education Reform Movement (Sahlberg, 2016) or New Public Management reforms in education (Gunter, Grimaldi, Hall, & Serpieri, 2016) are often used in education literature to capture this international phenomenon.

In this paper, we argue that, despite their globalizing dimension and neoliberal origins, the reception and evolution of data-intensive governance instruments needs to be seen as context-sensitive, contingent and path-dependent (Kauko, Rinne, & Takala, 2018; Maroy, Pons, & Dupuy, 2017). Policy instruments such as NLSAs and TBA have been globally adopted, but the uses (and the intensity of the uses) given to these instruments are arguably contingent to the specificities of the political and institutional settings where they are embedded. Governments might adopt NLSAs and TBA for different reasons, and it cannot be taken for granted that these instruments are chosen to promote similar policy changes, or that end up deepening market dynamics in education wherever they are enacted.

To test these premises, we have systematically reviewed a corpus of 158 papers focusing on the political economy of educational governance reforms in different world locations. We analyze our data through the lenses of a political sociology approach to policy instruments, which we combine with analytical premises deriving from historical institutionalism. In this respect, we expect that the politico-

administrative regimes to which countries adhere strategically mediate the variegated adoption and evolution of NLSAs and TBA in education. The paper is structured as follows. In the first two sections, we introduce our theoretical and methodological framework. In the sections that follow, we present our main results according to the main analytical axes of our approach: instruments choice, the evolution of instruments' uses, and the new constituencies and subjectivities generated by governance instruments. In the last section of the paper, we discuss our findings and conclude.

## A 'socio-historic' approach to policy instruments

The emerging demand for global skills in increasingly inter-dependent economies, the challenges generated by technological innovation, and the comparisons of educational systems deriving from international large-scale assessments are contributing to the expansion of similar education reforms globally (Verger et al., 2018). To a great extent, standardized evaluations and TBA have become central instruments of an education reform approach that situates learning achievement as a key driver of national success in an increasingly competitive economic environment (Rizvi & Lingard, 2010).

In this paper we acknowledge the importance of international competition, among other global drivers, for the spread of educational governance instruments. Nonetheless, we adopt a political sociology approach to policy instruments, in combination with a historical institutionalism premise, to more explicitly capture how meso-level factors strategically combine with global drivers in the production of more complex and multi-scalar education policy landscapes. This approach is well-suited to observe and systematize the diverging policy trajectories that global reforms and global policy models follow. To this purpose, we structure our theoretical framework according to two critical moments of educational policy change, first, the adoption of new policy instruments and, second, the evolving uses assigned to these instruments once they are being enacted.

### *Instrumentation: inquiring into the adoption of new policy instruments*

Policy instruments are central to both conceptualize and understand current public sector reforms and changing forms of governance (Le Galès, 2010). Although the adoption of new policy instruments has been conventionally conceptualized as a second order change (cf. Hall, 1993), or as a change of a mainly technical nature, policy instruments choice is a very political moment that does not always follow technocratic and pragmatic logics.

Instrumental choice is a moment with major potential implications for the future development of public systems. Many policy instruments create their own structures of opportunity in ways that were unforeseen when first adopted, and can generate broader political effects in governance structures and even in the main goals that policy systems are expected to pursue (Bezes, 2007; Kassim & Le Galès, 2010). For these reasons, political sociologists such as Lascoumes and Le Galès (2007) invite us to problematize the logic of instruments choice and conduct further empirical research on what they define as the moment of *instrumentation*.

The study of instrumentation focuses on policy-makers' discourses, interests and rationales when selecting new instruments, as well as on the range of economic, political and institutional contingencies that condition instrument selection (Capano & Lippi, 2017; Maroy et al., 2017, Peters, 2002). Economic factors, such as the level of economic development of a country, the international economic agreements that the country in question has signed, or periods of financial crisis, encourage or inhibit the selection of certain policy instruments and determine the financial feasibility of instrument options (Lenschow, Liefferink, & Veenman, 2005). On their part, political factors such as party politics and political ideologies are also expected to frame instruments' choice. Here, Le Galès (2010) refers to Bourdieu's metaphor on the right and the left sides of the State as a way to distinguish between 'a left democratic version [of governance] promoting negotiation, and more deliberative making of the general interest and a right mode of governance using indicators, standards and technical instruments to centralize and promote a more market-oriented society' (p. 143).

Politico-administrative regimes are also core mediating factors in instrumentation processes. Considering that data-intensive policy instruments aim at altering the way public services are administered and governed, prevailing administrative institutions will expectedly condition how these instruments are received, selected and enacted. Pollitt and colleagues (2007) have studied the path-dependent reception and evolution of New Public Management (NPM) instruments in different OECD territories from this perspective (see also Gunter et al., 2016). These authors start by noticing that, after decades of NPM promotion, 'OECD public administrations have become more efficient, more transparent and customer oriented, more flexible and more focused on performance' (OECD, 2005, p. 10 in Pollitt, 2007, p. 11). Nonetheless, they observe that NPM principles (such as decentralization, outcomes-based management, accountability and competition) have crystallized quite differently according to countries' public administration traditions. Thus, in countries with a more liberal organization of the State (such as the one prevailing in most Anglo-Saxon countries), NPM has adopted a more market-oriented form. In these countries, NPM has encouraged private sector participation in public services and more intense forms of competition between providers.

In contrast, in countries with a neo-Weberian State, which is the administrative model that mostly prevails in continental and northern Europe, NPM reforms have contributed to make services more citizen- and results-oriented, but not via the drastic promotion of market competition and choice. In these countries, the State adopts a proactive role as a facilitator of solutions to social problems and is eager to preserve the ideas of civil service and professionalism in public services (Pollitt, 2007).

Finally, within the so-called Napoleonic administrative tradition, which prevails mainly in Southern European countries and is characterized by centralized, hierarchical and uniform bureaucracies, NPM reforms 'have been tried, but with disappointing results' (Pollitt, 2007, p. 12). Components of NPM have been adopted disconnectedly and implemented unevenly. Furthermore, NPM reforms have met serious opposition from different flanks, which go from street-level to high-level bureaucrats (Kickert, 2007).

### *On the evolutionary dimension of policy instruments*

The evolution and future use(s) of policy instruments are conditioned by the previous instruments in place. Policy change often happens through the sedimentation of policy instruments. It operates as a *layering* process, in which policy change entails the addition of new instruments on top of existing ones; but it might also act as a *conversion* process in which instruments are adapted to new circumstances over time, and new uses and purposes are given to them (Thelen, 2004; Vetterlein & Moschella, 2014).

Policy instruments might change or evolve by inertia (rather than by design), reasons as to why their evolution and effects are often unpredictable (Mahoney, 2000). Instruments 'have impacts on their own, independent from the policy goals' (Le Galès, 2010, p. 151), or from 'the decisions that created them' (Kassim & Le Galès, 2010, p. 11). As a consequence, it is difficult to predict the form that any instrument will end up assuming, as well as their most direct effects (Bezes, 2007). To a great extent, actors' responses to new instruments might be more creative and diverging than policy-makers expect and, accordingly, both the responses and the effects produced by policy instruments 'depend on how the aims and purposes ascribed to them, and the meanings and representations they carry, are perceived, understood and responded to by key actors' (Skedsmo, 2011, p. 7).

Another important premise to understand the (unpredictable) evolution of policy instruments is that, once selected, instruments privilege certain actors and their interests (over others), and usually incentivize the generation of new constituencies. New constituencies are comprised of political and/or economic actors 'oriented towards developing, maintaining and expanding a specific instrumental model of governing' (Simons & Voß, 2018. p. 31) that have vested interests in the intensification of instrument uses. At the same time, however, new instruments might disserve other groups and, accordingly, trigger critical reactions and different forms of resistance.

To conclude, the perspective to policy instruments analysis presented here involves the systematic study of the sequence of contingencies (institutional, political, economic), events, and actions behind policy instruments' choice and their changing uses, and enquires into how and to what extent the selection and evolution of instruments contribute to advance substantive changes in the governance of public policy systems.

## Methods

The paper draws on the results of a systematic literature review of indexed publications focusing on the political economy of educational governance reforms. A systematic literature review (SLR) is a methodology oriented toward the synthesis of existing research on a particular topic and is characterized by the use of explicit and transparent methods of search and selection in order to reduce possible bias (Petticrew & Roberts, 2006).

Our research followed the conventional steps of SLRs as established by specialized literature. The main source considered was the SCOPUS database, although we also relied on recommendations from key informants with expertise on countries under-represented in the indexed literature.[1] After different screening processes, the final selection of papers considered in this review includes a total of 158 papers.

Information was systematized on the basis of country-specific extraction sheets. The reason to use country-specific (instead of paper-based) extraction sheets was that, frequently, reviewed papers deal with very specific components of the assessments and accountability framework of a country or tend to focus on policy changes in different periods of time. Aggregating fragmented pieces of literature into specific country-forms contributed to have a more comprehensive and accurate understanding of this complex education reform phenomenon.

## Instruments' choice: three rationales behind educational governance reforms

The reviewed literature allowed us to differentiate between three main policy trajectories in reforming educational governance, which echo the politico-administrative traditions developed by Pollitt (2007) and Pollitt and Bouckaert (2004) sketched above.

### *NPM marketizers: adopting TBA to expand market competition and choice*

In Anglo-Saxon countries, public sector reforms have been permeated by free-market policy principles and public choice theory. In the context of the global economic crisis of the 1970s, neoliberal ideas gained prominence in political and public policy agendas of countries such as the United States, England, New Zealand, but also Chile (see Falabella, 2015; Hursh, 2005). According to Pollit and Bouckaert (2004), these countries can be classified as 'core NPM states' or 'NPM marketizers' (p. 86), in the sense that the NPM toolkit was used strategically to advance the marketisation and the privatization of public services. In this context, education reforms adopted a market-driven approach and entailed the adoption of new governance instruments to steer 'at a distance' an increasingly complex and fragmented pool of educational providers.

The early adopters of NPM in education articulate a coherent and explicit theory of change on how TBA, parental choice and competition can trigger a sort of school improvement 'virtuous circle'. In these countries, national assessments and accountability measures have been enacted in combination with exogenous privatization policies (such as vouchers or other forms of public subsidies for private schools) with the purpose of stimulating market competition in education and empowering parents in their role as clients (Ball, 2008; Clarke, Gewirtz, & McLaughlin, 2000). NLSAs have been conceived as a pivotal policy instrument to collect data on schools' performance, to inform parental choice and to promote market accountability dynamics. Education reforms have been justified by a persistent discourse on public schooling failures and low-quality education in the public sector, usually attributed to burdensome bureaucratic rules and absence of incentives (Falabella, 2015; Hursh, 2005; Whitty & Wisby, 2016). The necessity of stricter surveillance mechanisms has been justified by a hostile discourse against teachers and teachers' unions. Thus, in a context of mistrust of public education, TBA and standards appear as a suitable policy solution to increase State control over public schools, teachers' work and the curriculum (Whetton, 2009).

In England and Chile, these changes in educational governance have enjoyed great stability over time, in great part, due to the fact that they are part of a profound and structural process of re-structuration of the state. In fact, once enacted, TBA systems have become increasingly complex and sophisticated, and their uses have been

incrementally intensified by different governments – regardless of their political ideology (Ball, 2008; Parcerisa & Falabella, 2017). In contrast, in the US, market reforms were tried in the 1980s but did not advance as quickly due to the complexities of the American political architecture (Klitgaard, 2007). In this country, school choice and TBA were not regulated as complementary policy instruments, at least at the Federal level, until the enactment of No Child Left Behind Act, passed in the year 2001 (Betebenner, Howe, & Foster, 2005). In the political process that led to the approval of the NCLB Act, discourses on competitiveness and choice were strategically combined with discourses about racial (and socioeconomic) equity and the reduction of achievement gaps (Hursh, 2005). The accountability pressures that came with NCLB opened the door to the conversion of underperforming public schools into charter schools, and widened up school exit and choice opportunities among families.

### *Neo-Weberian states: governance instruments travel to continental Europe*

In the last two decades, governance instruments have been widely adopted in continental Europe, including central and Nordic countries where a Neo-Weberian politico-administrative regime prevails. In Neo-Weberian states, external evaluations and new accountability instruments in education were not initially chosen to promote market competition, but as a way for the central State to guarantee quality standards in a context of highly decentralized education systems. In the 1980s and 1990s, Nordic countries like Denmark, Sweden or Norway went through profound decentralization processes which transferred numerous competences and responsibilities on education to local governments (Hatch, 2013; Moos, 2014). Decentralization in public services was adopted for subsidiarity reasons and as a way to make services more responsive to citizens' demands. However, decentralization came at the cost of central control and key stakeholders started questioning whether the national government had the necessary tools in place for monitoring the quality of education (Tveit, 2009).

Nonetheless, countries like Norway and Denmark did not react to quality control concerns until the first Programme for International Student Assessment (PISA) results were released in 2001. Disappointing results in PISA reinforced the social perception of a 'learning crisis' and opened an important window of opportunity for education reform advocates. As a consequence, most Nordic countries embraced an outcomes-based management approach to education and introduced more centralized (and standards-oriented) curricula (Møller & Skedsmo, 2013; Ydesen, 2013). They also adopted external evaluations as a way to both regain control over the curriculum and monitor the educational work of both local governments and schools more closely. Quality assurance schemes, national systems for school inspection and new education assessment units were also created (Helgoy & Homme, 2007).

In Central Europe, countries like Germany, Austria and The Netherlands have gone through a similar reform process, and have introduced standardized evaluation and accountability systems also as a way to regain control over curriculum delivery and academic results. Germany is one of the countries in Europe that has been more profoundly affected by unexpected low results in international large-scale assessments. In this country, the first PISA report generated a 'shock' that entailed the introduction of an output-oriented governance approach, which crystallized in three main interventions, namely the establishment of centralized monitoring of education standards, the strengthening of school autonomy, and the expansion of research-based

policy-making. These reforms were more intense in those German länders (or states) at the bottom of the inner-German PISA ranking (Niemann, Martens, & Teltemann, 2017). Nonetheless, despite the emphasis on accountability in education in the aftermath of PISA, the accountability systems adopted in Germany, as happened in most Neo-Weberian states, were predominantly low-stakes (Thiel, Schweizer, & Bellmann, 2017).

In a nutshell, in Central and Northern Europe, the adoption of new assessment and accountability practices is largely motivated by a sense of an international race for educational results. In many ways, in these countries, as stated by Browes and Altinyelken (2018) in relation to the Dutch case, accountability reforms 'can be understood as a rebalancing of the system, a "catch-up response" to the decentralization reforms that preceded them' (p. 13).

### *Napoleonic states: the partial and uneven adoption of governance reforms*

In Napoleonic states, which is the public administration model that prevails in Southern Europe, reforms to modernize public administration have been repeatedly tried, but have not always generated the expected changes (Gallego, 2003; Kickert, 2007). In these states, managerial education reforms have been adopted much later than in other education settings, without sufficient political backing, and not always following a comprehensive reform plan. In addition, once adopted, the implementation of accountability instruments has been uneven and highly conditioned by political contestation and economic junctures.

In Napoleonic states, most teachers have civil servant status and enjoy, *de facto,* high levels of autonomy. External evaluations have usually been adopted as a way to address an administrative sense of lack of control, and to encourage school improvement dynamics (Carvalho & Costa, 2017). To a great extent, new governance instruments and techniques have been adopted as a way to modernize the governance of the education system and to adhere to international norms and discourses on educational reform. However, these changes in the governance of education are not adopted as part of a cohesive and openly deliberated reform package, and their implementation has been often discontinued (Serpieri, Grimaldi, & Vatrella, 2015; Stamelos, Vassilopoulos, & Bartzakli, 2012). As we develop below, in Italy and in some Spanish regions, programs of merit-based pay, teacher/principal evaluation and school rankings have been repeatedly piloted, but discontinued after first attempts.

In the political discourse that predominates in the South-European region, new policy instruments, techniques and tools are usually attached to a rhetoric of 'quality assurance', but the *theory of change* of how these instruments are expected to generate quality gains is not always well articulated and explicit. Another characteristic of educational reforms in this region is the strong emphasis on 'school autonomy', which is usually translated into the promotion of a more hierarchical leadership style in schools (Dobbins & Christ, 2017). The professionalization and empowerment of the school principal figure represents a significant shift in countries with a long legacy of democratic and horizontal educational governance (with the principal being a *primus inter pares*, and many school decisions relying on community/families' participation) (Gairín Sallán, 2015). Democratic school governance (including the direct participation of families and teachers in core school decisions) in countries such as Portugal, Greece and Spain emerged as a reaction to decades of authoritarian regimes, which closely controlled the educational system for political reasons. Nonetheless, more recently, education reform advocates portray

democratic school governance as ineffective and promote the introduction of managerial changes and new leadership styles in schools (Veloso, Abrantes, & Craveiro, 2013; Verger & Curran, 2014).

## New and changing uses: the evolution of NLSAs and TBA systems

Besides the logic of instruments' choice, inquiring into the unfolding and evolutionary dimension of policy instruments is also necessary to understand the impact of instruments on educational governance. This section analyses the evolution of NLSAs and TBA systems, separately due to the fact that, from the perspective of their uses, NLSAs and TBA represent two analytically distinct varieties of instruments: respectively, instruments aimed at collecting information and instruments aimed at shaping behavior (Hood, 2007). Consequently, we should expect that both the pace and the nature of their evolution differ.

### *NLSAs: an ever-expanding instrument?*

Data collection instruments such as NLSAs tend to remain relatively stable over time and are rarely reversed. The main change experienced by NLSAs is related to their sophistication and expansion. With the passage of time, the frequency and scope of national assessments tend to increase. Changes in assessments are frequently the result of concerns on the reliability or relevance of the tests – for instance, when changes are introduced in order to ensure better alignment with curricular standards, or when the evaluated subjects are expanded over concerns about the narrowing of the curriculum. These preoccupations are often triggered by the increasing number of purposes given to (or expected to be served by) large-scale assessments, turning test validity into a matter of critical importance.

This intensification pattern holds across different contexts, but is particularly evident in core NPM countries, such as Chile and England. In Chile, during the late 1990s and early 2000s, as concerns over education quality became central to the education policy agenda, a series of changes were implemented to align national curriculum and assessment (Gysling, 2015). The tests were progressively reoriented towards the evaluation of a national curricular framework, in which cognitive skills of a superior order were measured through the introduction of new types of questions. Between 2005 and 2011, new grades and subjects were added to the national testing framework, and the test frequency was intensified so to allow student tracking over time (Bravo, 2011).

In other cases, the intensity of the testing framework is not necessarily altered, but new monitoring tools are devised – for instance, by adding new levels of data (dis)aggregation, creating new schools-classification systems, or developing new composite measures (indexes, typologies, etc.). The introduction of additional measures and more sophisticated tools is the logical consequence of the creation of specialized organizations (typically more independent evaluation agencies) whose main responsibility is to supervise and make use of a growing volume of collected data. These dynamics can be clearly observed in the English testing framework. While national assessments have remained relatively stable (in terms of number and frequency) since their introduction with the 1988 Education Reform Act (ERA), the number and sophistication of performance-related information has increased substantially – for instance with the adoption of the Pupil Achievement Trackers and

Performance and Assessment (Panda) reports. In addition, in 2010, a series of decisions contributed to the consolidation and expansion of the national testing framework. This included the introduction of a new test (Phonics Screening Check, age 6), the revision of the Early Years Foundation Stage Profile (age 5), and the establishment of a new performance measure (English Baccalaureate, which highlights the proportion of pupils of a given secondary school achieving high grades in the General Certificate of Secondary Education) (Bradbury, 2014; Mansell, 2011).

The US provides another example of the expansive nature of national assessments. The passing of NCLB in 2001 established an extensive testing framework by requiring states to test students in grades 3 through 8, and once in high school. Since then a number of techniques that tie student performance to teacher evaluations, as incentivized by the Race to the Top Act of 2011, have also emerged. Among them, value-added models feature prominently given its widespread use by several states and urban districts (Amrein-Beardsley & Holloway, 2017; Baker, Oluwole, & Green, 2013).

### *The ebb and flow of test-based accountability*

TBA systems experience a slower development than NLSAs, as they unfold gradually, and are more likely to undergo a rather uneven evolution, with some of their components being discarded after some time in place. Nonetheless, these dynamics play out quite differently according to the different administrative traditions sketched above.

In the case of *NPM marketizers*, stakes have primarily tended to increase and intensify. This occurs not only because reputational and market consequences tend to be reinforced over time, but also because administrative and bureaucratic stakes are likely to be added to the accountability system. In Chile, for instance, the national test was initially created with the aim of informing parents' choice and for curricular control purposes (Meckes & Carrasco, 2010, Falabella, 2015). However, soon after the first publication of test results in 1995, a salary bonus linked to schools' performance and a series of additional administrative sanctions and dispositions were adopted (Flórez, 2013). In the 2000s, school subsidies for low-income students became conditioned to the school compliance with State-defined learning goals, and low-ranked schools became more closely supervised, lost the capacity to autonomously administer public funding, and risk closure if they did not show signs of improvement. Remarkably, some of these new administrative dispositions were adopted by the center-left government as a means to correct market failures and reinforce the role of the State as a regulatory agent. At the same time, these same dispositions were supported by the Right because they perceived them as a way to reform education, but preserve and secure the market system at the same time (Parcerisa & Falabella, 2017).

Similarly, in England, different authors document a rise in the stakes associated to standardized tests during the late 1990s and early 2000s, under the New Labour government. Some of the consequences of the evaluations had a reputational nature – since they included the public posting of performance data – and were paralleled by the intensification of the school choice and competition agenda (Mansell, 2011; Muijs & Chapman, 2009). However, other consequences of a more material and administrative nature were added as well. New arrangements included 'light-touch' intervention policies for failing or low-performing schools as well as performance-

based pay schemes (Mansell, 2011; Whetton, 2009). These changes entailed a major departure from the accountability system put in place with the 1988 ERA, which relied essentially on market dynamics. However, a reform passed by the Coalition government formed by Conservatives and Liberal Democrats after the 2010 election meant a certain return to a market-based TBA system. New dispositions reduced the role and intervention capacity of the inspectorate in well-functioning schools, as those scoring as outstanding were deemed exempt from regular inspection (Mansell, 2011).

In so-called *Neo-Weberian states*, an intensification pattern of TBA measures is less clearly discernible. In fact, in some of these cases, it is even possible to detect a certain deceleration of the accountability agenda, resulting from the removal of the initially adopted market and reputational consequences. Although many Nordic countries have experimented with the publication of test results, most of them have finally stopped doing so, at least at the national level, due to concerns with the quality and reliability of the measures, but also because of the critical response of key stakeholders.

The case of Norway is illustrative of these logics of adjustment and reversal. In 2005, a Conservative-led coalition introduced a National Evaluation System that included for the first time a national standardized test. The Conservative government behind the reform expected to create pressure on schools through the combination of the public posting of test scores and the introduction of greater levels of school choice. Public posting of school results was however suspended in 2005, once the new center-left coalition came into power (Camphuijsen, Skedsmo, & Møller, 2018; Hatch, 2013). A similar pattern has been documented in Denmark. In 2005, a public order required schools to publish relevant information on their websites, including results from evaluation and teaching (also, the 2002 Act on Transparency had established similar requirements), and the Ministry of Education experimented with the publication of school rankings on its website. However, the publication of school results remained a particularly controversial issue, and the center-left coalition that came to power in 2011 suspended the publication of school league tables (Ydesen & Andreasen, 2014).

However, in both Denmark and Norway, the publication of school results re-emerged, even in the absence of national governmental action. In these two countries, even when the national government does not publish school scores, many local governments or the media do so, taking advantage of transparency rules in public administration. Similarly, in British Columbia (Canada), it is the Fraser Institute – an advocacy-oriented think tank explicitly committed to a deregulation and marketization agenda – that publishes the school report cards it produces through the media (Simmonds & Webb, 2013).

In *Napoleonic States*, TBA reforms appear to advance only through a trial-and-error logic and a back-and-forth dynamic characterized by frequent discontinuities. Test-based accountability remains relatively underdeveloped, and the raising of the stakes depends largely on the political orientation of the government in power. Attempts to introduce some form of principal evaluation in Italy illustrate these dynamics well. During the early 2000s, a series of principal-evaluation pilots were developed by a center-right government. Although evaluation was voluntary, the initiative faced high rates of rejection – partially as some assumed that the pilots were only the first step before the implementation of mandatory evaluation. The initiative was discontinued in 2006, as a center-left coalition took office. However, the election of a new center-right government in 2008, along with the appointment of a Minister of Education explicitly committed to meritocracy in education, put the principal

evaluation back to the policy agenda (Grimaldi & Serpieri, 2013). During the early 2010s, a new project establishing the voluntary evaluation of schools and teachers was proposed, giving rise in 2012 to an experimental evaluation model integrating the measurement of the school added value and the evaluation of principals (Serpieri et al., 2015). It was not until the approval of the education reform *La Buona Scuola*, passed in 2015 by a technocratic cabinet, that principal evaluation was consolidated as part of the new accountability framework (Montefiore, 2018).

Similar dynamics can be observed in Madrid, an autonomous region of Spain that, during the early 2000s introduced a series of market-based accountability measures, including the introduction of a standardized and census-based test combined with an increase in school choice freedom. With these policies in place, this region took an 'outlier' market trajectory that, apparently, did not fit well with the main characteristics of a Napoleonic administrative tradition. The adoption of these reforms owed much to the strong leadership, entrepreneurship and top-down government style of President Aguirre (2003–2012), who was strongly and personally committed to market freedoms in all types of sectors, including education. However, the reform was not resilient to Aguirre's resignation in 2012, and the new government – even when it was in hands of the same political party – abandoned some of the most emblematic market-accountability dispositions, including the publication of school results, and reduced the frequency of testing (Pagès & Prieto, 2018).

Finally, federal or highly decentralized states deserve separate consideration. In these cases, the progressive heightening of the stakes generally occurs at the sub-national level. Hence, even if the Federal government does not associate material or bureaucratic consequences to national assessments, local or state authorities can take advantage of these instruments being in place in order to adopt their own accountability measures (see for instance Gable & Lingard, 2015; Termes & Mentini, 2018).

## New constituencies: the emergence of economic and political subjects within educational governance reforms

### *Economic interests in testing*

Governance reforms in education have contributed to the emergence of a testing and measurement industry. The presence of this industry is bigger in those countries defined as NPM marketizers, where testing is more intensive and is usually attached to higher-stakes. In these countries, private companies, consultancies and research organizations such as Pearson, the Australian Council for Educational Research (ACER) or the Learning Bar benefit from substantive contracts with governments for the design, administration and/or data analysis of national assessments (Burch, 2006). These and other companies also sell school improvement services, lesson plans and/or educational platforms to those local governments and schools that aim at strengthening their performance (Hogan, Sellar, & Lingard, 2016). School improvement companies are very active in countries such as England and Chile, and many specialize in how to increase students' scores in external evaluations. School improvement is an important market niche in Chile, to a great extent, because schools can resort to public funding to hire these types of services (Osses, Bellei, & Valenzuela, 2015).

The expansion of economic interests in testing and measurement activities is key to understand the on-going spread of external evaluations and related accountability instruments. As the OECD (2013) acknowledges, the fact that 'standardized student assessment becomes a more profitable industry' means that 'companies have strong incentives to lobby for the expansion of student standardized assessment as an education policy, therefore, influencing the activities within the evaluation and assessment framework' (p. 51).

Nonetheless, public universities and research institutes are also involved in contracts for test design and data analysis, but their presence is relatively bigger in Continental Europe than in the Anglo-Saxon world. In many European countries, public agencies are centrally involved in testing-related activities. In England, universities and academic centers are also involved in test development, but smaller academic suppliers have been largely withdrawing from testing arrangements due to the inclusion of stricter conditions in the contracts, such as penalties for not meeting the deadlines (Whetton, 2009).

Overall, the increasing involvement of both public and increasingly private groups within education testing activities (both as third-party producers and as parties supporting policy implementation) explains why testing and TBA instruments expand not only territorially, but also toward new areas of educational activity and education levels. When the political and economic interests of these groups are strong, there are more reasons to expect that these policy instruments will endure in time, independently of their effectiveness (see Dale, 2018). In other words, the political and economic positioning of the groups that administer governance instruments – or deliver related services – within the education policy field is key to explain policy continuity and lock-in effects, even when the effectiveness of intensive uses of standardized testing is increasingly questioned by academic evidence.

### *Spreading like wildfire, but meeting firewalls*

Standards, assessments and accountability are known for 'spreading like wildfire', among other reasons, because they are inexpensive to fund in comparison to other policy alternatives (Hargreaves & Shirley, 2009). The fact that the adoption of policy instruments such as NLSAs or TBA results from debates of a technical (rather than political) nature also makes the articulation of social responses difficult. National assessments tend to generate less resistance because they are usually adopted as non-intrusive data-gathering instruments, and their ultimate effect on teachers and schools is not evident since the outset. However, under certain circumstances, these instruments also generate critical reactions.

TBA is the most contentious governance instrument, especially when associated with high-stakes outcomes. High-stakes accountability usually triggers passionate debates between its supporters and detractors, and is also a motive of collective action (Pizmony-Levy & Woolsey, 2017). Nonetheless, responses to TBA are not always led by teachers' unions (TUs). In fact, in the light of existing literature, TUs responses to accountability reforms cannot be taken for granted. On the contrary, accountability usually places TUs in difficult political dilemmas, a major reason why their responses are variegated and not always as defiant as could be expected. For example, in Chile the main TU (*Colegio de Profesores*) did consent – and participated in the definition of – new forms of teacher evaluation policies. The union saw these policies as a lesser evil in a context of profound marketization and pauperization of teachers' work, and

as a *de facto* opportunity to raise teachers' salaries (Gindin & Finger, 2013; Vaillant, 2005).

In the context of the US, the main unions were very cautious regarding TBA under the NCLB Act and, despite their general dissatisfaction with the reform, they could not agree on a unitary response (Hursh, 2005). Furthermore, given the fact that NCLB had such a strong equity discursive frame, the leadership of the unions wanted to avoid being seen as insensitive to children learning issues or to existing learning gaps in front of society. The teachers' critique to standardized testing is particularly challenging in the US because, first, the societal trust in teachers is relatively low and, second, standardized testing is sound with values that are deeply rooted in American society such as meritocracy, achievement and effort (Au, 2016).

In Nordic European countries, TUs opposition to TBA is not only explained by ideological reasons, but by normative and professional understandings of how the teaching profession should be regulated. In Norway, teachers adopted a critical attitude towards NLSAs policies because they felt their professional autonomy and judgment capacity was being challenged, although they did not articulate a confrontational campaign against the national assessment first implemented in 2004. Their critical position against this assessment was not only ideologically motivated, but emphasized the poor design and quality of the test (Tveit, 2009). In contrast, in Sweden, NLSAs did not generate so much controversy. There, teachers did not see the national assessment as an external instrument challenging their autonomy, but as a useful tool to support their professional development and to establish quality standards in education (Helgoy & Homme, 2007).

A more oppositional approach among TUs appears in Southern Europe. In Italy, the strong bargaining capacity of the TUs is crucial to understand the discontinuities and changes experienced by many managerial programs and tools (Barzanò & Grimaldi, 2013; Grimaldi & Serpieri, 2013). In the Portuguese context, national evaluations were perceived as repressive instruments with reminiscences to the dictatorial period, and generated fierce resistance by well-organized TUs (Veloso et al., 2013). In general, in Southern Europe, TUs, but also other local stakeholders and parent associations, are especially belligerent against accountability instruments that more directly challenge teachers' autonomy and the democratic governance of schools.

The adoption of TBA also generates critical responses from social actors other than TUs, including students' movements, critical scholars, pedagogical associations and, especially, parents. In fact, these social responses tend to emerge in contexts where TUs adopt an ambiguous or passive position in front of standardized testing and accountability reforms. For example, in Norway, the government approved a moratorium in standardized testing after students' boycotts were carried out in the main cities of the country. This boycott was supported by teachers, but only covertly (Helgoy & Homme, 2007).

In the US, the rapid intensification of standardized testing has triggered the emergence of an ideologically transversal opt-out movement that crystallizes in organizations such as *FairTest* or *United Opt Out* (Dobrick, 2014; Pizmony-Levy & Woolsey, 2017). In states like New York, families have massively followed calls to boycott the test and, in response to these actions, the government has introduced improvements in the test design and administration. The political influence of this movement also manifests in the fact that 'opt-out parents' have been elected in different district school boards (Wang, 2017).

More recently, similar social movements that boycott national tests have emerged in other countries, such as Spain (see Saura et al., 2017), and Chile (Campos-Martínez & Guerrero-Morales, 2016; Pino-Yancovic, Oyarzún-Vargas, & Salinas-Barrios, 2016). These social movements, which tend to be led by middle-class families, articulate a very sophisticated narrative regarding the non-desired educational effects of standardized testing, and are particularly skillful in managing social media and performing innovative collective actions.

## Discussion and conclusions

In the last decades, OECD and middle-income countries have given greater salience to external evaluations, targets, standards and accountability in the governance of their educational systems. Policy communities share, internationally, a similar discourse on evaluation and accountability and, at the regulatory level, we have observed that countries are almost unanimously adopting NLSAs and TBA instruments when reforming their educational systems. However, both the instrumentation process (i.e., the political process through which policy instruments are being adopted) and the uses that governments are giving to policy instruments are not converging so clearly cross-nationally.

Institutional legacies strategically mediate the adoption of education governance instruments. It is mainly in *Anglo-Saxon countries*, with a liberal organization of the State, where these instruments have been adopted with a more obvious pro-market purpose. In these countries, evaluation and accountability instruments are explicitly used to promote school competition and choice, and are more clearly attached to school rankings and merit-based pay formulas. With the passage of time, accountability systems have become more complex and sophisticated, and their stakes higher. The more established political forces (Liberal, Conservative and Labour parties) agree on the central role of testing and TBA in educational reforms, and on the main uses that should be given to these instruments. Accordingly, the public posting of schools' results is less likely to be questioned and, in fact, the production of school rankings has frequently been promoted by center-left parties, in an attempt to make public services more transparent and democratize school choice. A testing industry has emerged more strongly in the context of NPM marketizers, and TUs responses to these policies has been rather timid.

In contrast, in *neo-Weberian states,* evaluation and accountability instruments have been adopted following a quality assurance rationale, in an education policy landscape characterized by high levels of decentralization. The most important political parties agree on the adoption of assessments and accountability systems, but not always on their uses, with the right more inclined to produce rankings and promote market competition than the left.[2] Accordingly, the uses of governance instruments vary in different legislative terms, but are also highly contingent to local politics.

Finally, in *Napoleonic states,* accountability has been adopted at the regulatory level and to comply with international norms and discourses on educational governance, but its enactment is very uneven and frequently remains in a latent or incipient stage. In these countries, the advancing of governance reforms has been more clearly conditioned by political and economic junctures. The political consensus (between the right and the left) around external assessments and accountability is not as evident as in the previous cases, and TUs have confronted more directly TBA and

other managerial reforms, usually through industrial action. Overall, in both Neo-Weberian and Southern European countries, governance instruments are not always adopted as a synonym of a market competition agenda, but tend to generate increasing performative pressures and unrest among key education stakeholders.

The main political forces advocating the adoption of national assessments and accountability instruments also differ across contexts. According to the literature reviewed, in mainland Europe, the OECD is at the center of a transnational *instrument constituency* (cf. Béland & Howlett, 2016) that effectively advocates school autonomy with accountability as a superior form of educational governance. There, the OECD, mainly via PISA, but also through other policy mechanisms, has strongly triggered national educational debates that have derived into significant changes in the education sector. In contrast, in core NPM states, the origin of educational governance reforms is more endogenous and, accordingly, the role of domestic policy entrepreneurs, consultants and think tanks is more often reported than the role of international organizations.

As a note of caution, the policy trajectories that we have identified are not exhaustive. They do not reflect all possible reform manifestations, but those that are more widely represented in existing literature. As more research on the topic is being produced in different world regions, the more feasible it will be to complement and widen our categories. Furthermore, although the policy trajectories are informed by specific public administration regimes, these same administrative regimes cannot be taken as a 'kind of unchanging bedrock' (Pollitt & Bouckaert, 2011, p. 48, cited in Gunter et al., 2016, p. 16). Federal countries are particularly difficult to classify from the administrative regimes' perspective, because their own states might lead to diverging reform trajectories.

### *On the cumulative nature of educational governance reforms*

The educational transformations that we have analyzed in this paper have a fragmented and cumulative nature in the sense that, in most cases, policy change advances as the result of the sedimentation and layering of different instruments, techniques and tools that are not necessarily articulated in a predefined reform program. With the passage of time, instruments such as accountability systems gain autonomy from their promoters and, accordingly, adopt functions and generate effects that were not initially foreseen.

Furthermore, with the exception of NPM marketizers such as Chile and England, NLSAs and TBA are generally adopted in different points in time and, in fact, as the result of differentiated debates, and in response to different demands. The implementation of a NLSA system generally precedes the adoption of TBA measures (since some form of census-based assessment is a necessary condition for TBA to happen) and, with the passage of time, economic, bureaucratic or reputational consequences tend to be attached to the assessments. Test consequences are usually State-mandated, but many other consequences emerge independently from governmental policies and intentions (e.g., the role of the media or philanthropic organizations in the dissemination of test scores). Overall, the cumulative nature of governance reforms makes it particularly necessary to get an accurate, sequential and diachronic understanding of the development of this model of educational change.

Our results also suggest that policy instrumentation is not a fixed moment in time, and goes through recurrent back-and-forth dynamics. These dynamics are, on

occasions, the result of changing political and economic junctures, but also the result of effective collective actions (such as those organized by TUs or the opt-out movement). Nonetheless, even if collective action might alter the uses (or the design) of policy instruments, it is less likely to challenge the very existence of these instruments.

The layering process through which national assessments and TBA evolve explains why it is so difficult to challenge the presence of these instruments in educational systems. Key stakeholders do not foresee many of the changes that these instruments involve from the beginning, and they only start reacting to them once their uses have intensified and been routinized. Furthermore, our findings reflect that there are several factors that have a lock-in effect on the use of external evaluations and accountability mechanisms. These include the emergence of new constituencies and economic interests around testing activities; the international education environment that, among other things, favors an international education race for constantly-improving learning results; and the broad political consensus generated around data-intensive governance instruments.

The malleable nature of governance instruments (i.e. they might serve to address learning gaps and strengthen the democratic control of education, but also to promote school choice and competition) contributes to their widespread and politically transversal adoption, as well as to their incremental evolution. From the perspective of instrumentation, external evaluations and accountability are appealing and convenient choices. The seductive power of these instruments relies on the fact that they contribute to transform complex and multi-dimensional educational realities into numerical categories, and to construct the perception that deep educational problems (such as inequalities or quality issues) can be addressed by setting up predefined patterns of conduct, measuring actors' performance, and distributing incentives accordingly (Falabella, 2018-forthcoming, Barbana, Dupriez, Dumay, 2014).

For all these reasons, data-intensive policy instruments continue expanding globally and are gaining centrality in the governance of education systems. This globalizing phenomenon urges us to retrieve new sources of evidence on the intended and unintended effects of educational governance reforms in schools' micro-dynamics. Future research on the topic could explore the multiple appropriations of governance instruments at the school level, and analyze under what conditions these instruments generate more instrumental or expressive responses among teachers and principals. This type of research, which is particularly underdeveloped in relation to soft-accountability systems, could contribute to inform better policy decisions in the future. By unraveling the political connotations and implications of policy instruments, research can also contribute to promote more democratic and informed debates about educational change in the governance era.

## Notes

[1] The search terms can be consulted in the Appendix.

[2] Although the reviewed literature also shows that social-democratic governments tend to be more belligerent with certain TBA uses when in opposition than when in power (see for instance Solhaug, 2011).

## References


Amrein-Beardsley, A., & Holloway, J. (2017). Value-added models for teacher evaluation and accountability: Commonsense assumptions. *Educational Policy*. http://doi.org/10.1177/0895904817719519

Au, W. (2016). Meritocracy 2.0: High-stakes, standardized testing as a racial project of neoliberal multiculturalism. *Educational Policy*, *30*(1), 39–62.

Baker, B. D., Oluwole, J. O., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the Race-to-the-Top era. *Education Policy Analysis Archives*, *21*(5). http://doi.org/10.14507/epaa.v21n5.2013

Ball, S. J. (2008). The legacy of ERA, privatization and the policy ratchet. *Educational Management Administration & Leadership*, *36*(2), 185–199. http://doi.org/10.1177/1741143207087772

Ball, S. J., Junemann, C., & Santori, D. (2017). *Edu.net. Globalisation and education policy mobility*. London: Routledge. http://doi.org/10.4324/9781315630717

Barbana, S., Dumay, X., & Dupriez, V. (2014). Perceptions et usages des instruments d'accountability. Enquête exploratoire dans l'enseignement secondaire en Belgique francophone. [Perceptions and uses of accountability instruments. Exploratory survey in secondary education in French-speaking Belgium.] *Éducation compare*, *12*, 21-44.

Barzanò, G., & Grimaldi, E. (2013). Discourses of merit. The hot potato of teacher evaluation in Italy. *Journal of Education Policy*, *28*(6), 767–791. http://doi.org/10.1080/02680939.2013.774439

Béland, D., & Howlett, M. (2016). How solutions chase problems: Instrument constituencies in the policy process. *Governance*, *29*(3), 393–409.

Betebenner, D. W., Howe, K. R., & Foster, S. S. (2005). On school choice and test-based accountability. *Education Policy Analysis Archives*, *13*(41). http://doi.org/10.14507/epaa.v13n41.2005

Bezes, P. (2007). The hidden politics of administrative reform: Cutting French civil service wages with a low-profile instrument. *Governance*, *20*(1), 23–56. http://doi.org/10.1111/j.1468-0491.2007.00343.x

Bradbury, A. (2014). 'Slimmed down' assessment or increased accountability? Teachers, elections and UK government assessment policy. *Oxford Review of Education*, *40*(5), 610–627. http://doi.org/10.1080/03054985.2014.963038

Bravo, J. (2011). SIMCE: Pasado, presente y futuro del sistema nacional de evaluación [SIMCE: Past, present and future of the national system of evaluation]. *Estudios Públicos*, 123, 189–211.

Browes, N., & Altinyelken, H. K. (2018, September). The evolution of test-based accountability in the autonomous Dutch system: A policy instruments approach. Paper presented at the European Conference on Educational Research (ECER), Bolzano, Italy.

Burch, P. E. (2006). The new educational privatization: Educational contracting and high stakes accountability. *Teachers College Record*, *108*(12), 2582–2610.

Campos-Martínez, J., & Guerrero-Morales, P. (2016). Efectos indeseados de la medición de la calidad educativa en chile. La respuesta de la sociedad civil [Undesired effects of quality measuring in Chile. The civil society response]. *Cadernos CEDES*, *36*(100), 355–374. http://doi.org/10.1590/cc0101-32622016171351

Camphuijsen, M. Skedsmo, G., & Møller, J. (2018, September). School autonomy with accountability as a global education reform: Its adoption and re-contextualization in the Norwegian context. Paper presented at the European Conference on Educational Research (ECER), Bolzano, Italy.

Capano, G., & Lippi, A. (2017). How policy instruments are chosen: Patterns of decision makers' choices. *Policy Sciences*, *50*(2), 269–293. http://doi.org/10.1007/s11077-016-9267-8

Carvalho, L. M., & Costa, E. (2017). Avaliação externa das escolas em Portugal: atores, conhecimentos, modos de regulação [External evaluation of schools in Portugal: Actors, knowledges and regulation modes]. *Revista Brasileira de Política e Administração Da Educação*, *33*(3), 685–705. http://doi.org/10.21573/vol33n32017.79302

Clarke, J., Gewirtz, S., & McLaughlin, E. (2000). *New managerialism, new welfare*. Thousand Oaks: SAGE Publications.

Dale, R. (2018). Global education policy: Creating different constituencies of interest and different modes of valorisation. In A. Verger, H. K. Altinyelken, & M. Novelli (Eds.), *Global education policy and international development: New agendas, issues and policies* (pp. 289–298). New York: Bloomsbury.

Dobbins, M., & Christ, C. (2017). Do they matter in education politics? The influence of political parties and teacher unions on school governance reforms in Spain. *Journal of Education Policy*. http://doi.org/10.1080/02680939.2017.1406153

Dobrick, A. (2014). Poverty and pretense: Good intentions and misguided educational reform from No Child Left Behind through Race to the Top. In E. M. Zamani-Gallaher (Ed.), *The Obama administration and educational reform* (pp. 27–44). Bingley: Emerald Group Publishing Limited. http://doi.org/10.1108/S1479-358X20130000010002

Falabella, A. (2015). El mercado escolar en Chile y el surgimiento de la nueva gestión pública: el tejido de la política entre la dictadura neoliberal y los gobiernos de la centroizquierda. [The school market in Chile and the rise of new public management: the political intersection between the neoliberal dictatorship and center-left governments.] *Educação & Sociedade*, *36*(132), 699–722. http://doi.org/10.1590/ES0101-73302015152420

Falabella, A. (2018-forthcoming). The seduction of hyper-surveillance: Standards, testing and accountability policies in Chilean schools. *Educational Administration Quarterly.*

Flórez, M. T. (2013). *Análisis crítico de la validez del sistema de medición de la calidad de la educación (SIMCE)* [Critical analysis of the validity of the system of measurement of education quality (SIMCE)]. Oxford University Centre for Educational Assessment: Oxford.

Gable, A., & Lingard, B. (2015). NAPLAN data: A new policy assemblage and mode of governance in Australian schooling. *Policy studies*, 37(6), 568–582. https://doi.org/10.1080/01442872.2015.1115830

Gairín Sallán, J. (2015). Autonomy and school management in the Spanish context. *Educational, Cultural and Psychological Studies*, 11, 103–117. http://doi.org/10.7358/ecps-2015-011-gair

Gallego, R. (2003). Public management policy making in Spain, 1982–1996: Policy entrepreneurship and (in) opportunity windows. *International Public Management Journal, 6*(3), 283–307.

Gindin, J., & Finger, L. (2013). *Promoting education quality: The role of teachers' unions in Latin America*. (Paper commissioned for the EFA Global Monitoring Report 2013/2014, Teaching and learning: Achieving quality for all). Retrieved from scholar.harvard.edu/ files/lesliefinger/files/unesco_paper.pdf

Grimaldi, E., & Serpieri, R. (2013). Jigsawing education evaluation. Pieces from the Italian New Public Management puzzle. *Journal of Educational Administration and History*, *45*(4), 306–335. http://doi.org/10.1080/00220620.2013.822350

Gunter, H. M., Grimaldi, E., Hall, D., & Serpieri, R. (Eds.). (2016). *New public management and the reform of education: European lessons for policy and practice*. New York/London: Routledge.

Gysling, J. (2015). The historical development of educational assessment in Chile: 1810–2014. Assessment in education: *Principles, Policy & Practice*, *23*(1), 8–25. http://doi.org/10.1080/0969594x.2015.1046812

Hall, P. A. (1993). Policy paradigms, social learning, and the State: The case of economic policymaking in Britain. *Comparative Politics*, *25*(3), 275–296. http://doi.org/10.2307/422246

Hargreaves, A., & Shirley, D. (2009). The *fourth way. The inspiring future of educational change.* Thousand Oaks: Corwin.

Hatch, T. (2013). Beneath the surface of accountability: Answerability, responsibility and capacity-building in recent education reforms in Norway. *Journal of Educational Change*, *14*(2), 113–138. http://doi.org/10.1007/s10833-012-9206-1

Helgøy, I., & Homme, A. (2007). Towards a new professionalism in school? A comparative study of teacher autonomy in Norway and Sweden. *European Educational Research Journal*, *6*(3), 232–249. http://doi.org/10.2304/eerj.2007.6.3.232

Hogan, A., Sellar, S., & Lingard, B. (2016). Commercialising comparison: Pearson puts the TLC in soft capitalism. *Journal of Education Policy*, *31*(3), 243–258. http://doi.org/10.1080/02680939.2015.1112922

Hood, C. (2007). Intellectual obsolescence and intellectual makeovers: Reflections on the tools of government after two decades. *Governance*, *20*(1), 127–144. http://doi.org/10.1111/j.1468-0491.2007.00347.x

Hursh, D. (2005). The growth of high-stakes testing in the USA: Accountability, markets and the decline in educational equality. *British Educational Research Journal*, *31*(5), 605–622. http://doi.org/10.1080/01411920500240767

Kassim, H., & Le Galès, P. (2010). Exploring governance in a multi-level polity: A policy instruments approach. *West European Politics*, *33*(1), 1–21. http://doi.org/10.1080/01402380903354031

Kauko, J., Rinne, R., & Takala, T. (Eds.). (2018). *Politics of quality in education. A comparative study of Brazil, China, and Russia.* Abingdon/New York: Routledge. http://doi.org/10.4324/9780203712306

Kickert, W. (2007). Public management reforms in countries with a Napoleonic state model: France, Italy and Spain. In C. Pollitt, S. V. Thiel, V. Homburg, & S. Van Thiel (Eds.), *New public management in Europe* (pp. 26–51). London: Palgrave Macmillan UK. http://doi.org/10.1057/9780230625365_3

Klitgaard, M. B. (2007). Do welfare state regimes determine public sector reforms? Choice reforms in American, Swedish and German schools. *Scandinavian Political Studies*, *30*(4), 444–468. http://doi.org/10.1111/j.1467-9477.2007.00188.x

Lascoumes, P., & Le Galès, P. (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. http://doi.org/10.1111/j.1468-0491.2007.00342.x

Le Galès, P. (2010). Policy instruments and governance. In M. Bevir (Ed.), *Handbook of governance* (pp. 142–159). London/Thousand Oaks: Sage.

Lenschow, A., Liefferink, D., & Veenman, S. (2005). When the birds sing. A framework for analysing domestic factors behind policy convergence. *Journal of European Public Policy*, *12*(5), 797–816. http://doi.org/10.1080/13501760500161373

Mahoney, J. (2000). Path dependence in historical sociology. *Theory and Society*, *29*(4), 507–548.

Mansell, W. (2011). Improving exam results, but to what end? The limitations of New Labour's control mechanism for schools: Assessment-based accountability. *Journal of Educational Administration and History*, *43*(4), 291–308. http://doi.org/10.1080/00220620.2011.606896

Maroy, C., Pons, X., & Dupuy, C. (2017). Vernacular globalisations: Neo-statist accountability policies in France and Quebec education. *Journal of Education Policy*, *32*(1), 100–122.

Meckes, L., & Carrasco, R. (2010). Two decades of SIMCE: An overview of the national assessment system in Chile. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 233–248. http://doi.org/10.1080/09695941003696214

Møller, J., & Skedsmo, G. (2013). Modernising education: New public management reform in the Norwegian education system. *Journal of Educational Administration and History*, *45*(4), 336–353. https://doi.org/10.1080/00220620.2013.822353

Montefiore, G. (2018). The good school reform. A double analysis of the last Italian NPM Education reform in its development from government idea to law and in its rhetoric. Unpublished manuscript, Universitat Autònoma de Barcelona, Spain.

Moos, L. (2014). Educational governance in Denmark. *Leadership and Policy in Schools*, *13*(4), 424–443. http://doi.org/10.1080/15700763.2014.945655

Muijs, D., & Chapman, C. (2009). Accountability for improvement: Rhetoric or reality? In C. Chapman, & H. M. Gunter (Eds.), *Radical reform. Perspectives on an era of educational change* (pp. 28–41). Abingdon/New York: Routledge.

Niemann, D., Martens, K., & Teltemann, J. (2017). PISA and its consequences: Shaping education policies through international comparisons. *European Journal of Education*, *52*(2), 175–183. http://doi.org/10.1111/ejed.12220

OECD. (2013). *Synergies for better learning: An international perspective on evaluation and assessment*. Paris: OECD. Retrieved from: http://www.oecd.org/education/school/synergies-for-better-learning.htm

Osses, A., Bellei, C., & Valenzuela, J. P. (2015). External technical support for school improvement: Critical issues from the Chilean experience. *Journal of Educational Administration and History*, *47*(3), 272–293. http://doi.org/10.1080/00220620.2015.1038699

Pagès, M., & Prieto, M. (2018, September). SAWA policies in a context of school choice: Evidence from Spain. Paper presented at the European Conference on Educational Research (ECER), Bolzano, Italy.

Parcerisa, L., & Falabella, A. (2017). La consolidación del Estado evaluador a través de políticas de rendición de cuentas: Trayectoria, producción y tensiones en el sistema educativo. [The consolidation of the evaluative state through accountability policies: Trajectory, enactment and tensions in the Chilean education system.] *Education Policy Analysis Archives*, *25*(89). http://doi.org/10.14507/epaa.25.3177

Peters, G. (2002). The politics of tool choice. In L. M. Salamon (Ed.), with O. V. Elliott, *The tools of government. A guide to the new governance* (pp. 552–564). Oxford: Oxford University Press.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford, UK: Blackwell Publishing.

Pino-Yancovic, M., Oyarzún-Vargas, G., & Salinas-Barrios, I. (2016). Crítica a la rendición de cuentas: narrativa de resistencia al sistema de evaluación en Chile. [A critical approach to accountability: a narrative of resistance against the evaluation system in Chile.] *Cadernos CEDES*, *36*(100), 337–354. http://doi.org/10.1590/cc0101-32622016171362

Pizmony-Levy, O., & Woolsey, A. (2017). Politics of education and teachers' support for high-stakes teacher accountability policies. *Educational Policy Analysis Archives*, *25*(87). http://doi.org/10.14507/epaa.25.2892

Pollitt, C. (2007). Convergence or divergence: What has been happening in Europe? In C. Pollitt, S. V. Thiel, V. Homburg, & S. Van Thiel (Eds.), *New public management in Europe* (pp. 10–25). London: Palgrave Macmillan UK. http://doi.org/10.1057/9780230625365_2

Pollitt, C., & Bouckaert, G. (2004*). Public management reform. A comparative analysis* (2nd ed). Oxford. Oxford University Press.

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. http://doi.org/10.1086/698326

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

Sahlberg, P. (2016). The global educational reform movement and its impact on schooling. In K. Mundy, A. Green, B. Lingard, & A. Verger (Eds.), *The handbook of global education policy* (pp. 128–144). West Sussex: Wiley-Blackwell.

Saura, G., Muñoz-Moreno, J. L., Luengo-Navas, J., & Martos, J. M. (2017). Protestando en Twitter: Ciudadanía y empoderamiento desde la educación pública. [Protesting on Twitter: Citizenship and empowerment from public education]. *Comunicar: Revista Científica Iberoamericana de Comunicación y Educación*, *25*(53), 39–48. http://doi.org/10.3916/C53-2017-04

Scott, C. (2000). Accountability in the regulatory state. *Journal of Law and Society*, 27(1), 38–60. https://doi.org/10.1111/1467-6478.00146

Serpieri, R., Grimaldi, E., & Vatrella, S. (2015). School evaluation and consultancy in Italy. Sliding doors towards privatisation? *Journal of Educational Administration and History*, *47*(3), 294–314. http://doi.org/10.1080/00220620.2015.1038695

Simmonds, M., & Webb, P. T. (2013). Accountability synopticism : How a think tank and the media developed a quasi- market for school choice in British Columbia. *International Education Journal: Comparative Perspectives*, *12*(2), 21–41. Retrieved from https://openjournals.library.sydney.edu.au/index.php/IEJ/article/view/7454

Simons, A., & Voß, J.P. (2018). The concept of instrument constituencies: Accounting for dynamics and practices of knowing governance. *Policy and Society*, *37*(1), 14–35, doi: 10.1080/14494035.2017.1375248

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

Solhaug, T. (2011). New public management in educational reform in Norway. *Policy Futures in Education*, *9*(2), 267–279. http://doi.org/10.2304/pfie.2011.9.2.267

Stamelos, G., Vassilopoulos, A., & Bartzakli, M. (2012). Understanding the difficulties of implementation of a teachers' evaluation system in Greek primary education: From national past to European influences. *European Educational Research Journal*, *11*(4), 545–557. http://doi.org/10.2304/eerj.2012.11.4.545

Teltemann, J., & Jude, N. (2018, July). New accountability schemes? Assessing trends in educational assessment and accountability procedures in OECD countries. Paper presented at the XIX International Sociological Association (ISA) World Conference, Toronto, Canada.

Termes, A., & Mentini, L. (2018, September). From low to high stakes: Ideational and material drivers behind test-based accountability reforms in Minas Gerais, Brazil. Paper presented at the European Conference on Educational Research (ECER), Bolzano, Italy.

Thelen, K. (2004). How institutions evolve. In J. Mahoney, & D. Rueschemeyer (Eds.), *Comparative historical analysis in the social sciences* (pp. 208–240). Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9780511803963.007

Thiel, C., Schweizer, S., & Bellmann, J. (2017). Rethinking side effects of accountability in education: Insights from a multiple methods study in four German school systems. *Education Policy Analysis Archives*, *25*(93). http://doi.org/10.14507/epaa.25.2662

Tveit, S. (2009). Educational assessment in Norway – A time of change. In C. Wyatt-Smith, & J. J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice* (pp. 227–243). Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-1-4020-9964-9 12

Vaillant, D. (2005). *Education reforms and teachers' unions: Avenues for action* (IIEP Fundamentals of educational planning 82). Paris, France: UNESCO–International Institute for Educational Planning. Retrieved from unesdoc.unesco.org/images/ 0014/001410/141028e.pdf

Veloso, L., Abrantes, P., & Craveiro, D. (2013). The Portuguese schools' evaluation programme: A sociological approach to the participation of social actors. *Evaluation*, *19*(2), 110–125. http://doi.org/10.1177/1356389013485616

Verger, A., & Curran, M. (2014). New public management as a global education policy: Its adoption and re-contextualization in a Southern European setting. *Critical Studies in Education*, *55*(3), 253–271. http://doi.org/10.1080/17508487.2014.913531

Verger, A., Parcerisa, L., & Fontdevila, C. (2018). The growth and spread of large-scale assessments and test-based accountabilities: a political sociology of global education reforms. *Educational Review, 71(1),* 1-26. https://doi.org/10.1080/00131911.2019.1522045

Vetterlein, A., & Moschella, M. (2014). International organizations and organizational fields: Explaining policy change in the IMF. *European Political Science Review*, *6*(01), 143–165. http://doi.org/10.1017/S175577391200029X

Wang, Y. (2017). The social networks and paradoxes of the opt-out movement amid the Common Core State Standards implementation: The case of New York. *Education Policy Analysis Archives, 25*(1), 1–27.

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

Whitty, G., & Wisby, E. (2016). Education in England – A testbed for network governance? *Oxford Review of Education*, *42*(3), 316–329. http://doi.org/10.1080/03054985.2016.1184873

Wilkins, A., & Olmedo, A. (Eds.). (2019). *Education governance and social theory: Interdisciplinary approaches to research*. London: Bloomsbury Publishing.

Ydesen, C. (2013). Educational testing as an accountability measure: Drawing on twentieth-century Danish history of education experiences. *Paedagogica Historica*, *49*(5), 716–733. http://doi.org/10.1080/00309230.2013.815235

Ydesen, C., & Andreasen, K. E. (2014). Accountability practices in the history of Danish primary public education from the 1660s to the present. *Education Policy Analysis Archives*, *22*(120). http://doi.org/10.14507/epaa.v22.1618

## Appendix: List of terms included in the search protocol

TITLE-ABSTRACT-KEYWORDS (("accountability" OR (("high-stakes" OR "low-stakes") W/2 "test*") OR "managerial*" OR "incentive" OR "sanction" OR "reward" OR "bonus" OR "performance-based pay" OR "pay-per-performance" OR "merit-based pay" OR "teacher evaluation" OR "feedback" OR "school evaluation" OR "school self-evaluation" OR "school inspection" OR "standardi?ed test*") AND ("test*" OR "learning outcomes" OR "standard" OR "evaluat*" OR "rank*" OR "benchmark*" OR "assess*" OR "result" OR "grad*" OR ("performance" W/2 ("teacher" OR "student" OR "school"))))

AND TITLE-ABSTRACT-KEYWORDS ("International organi?ation" OR "IO" OR "NGOs" OR "unions" OR "teachers organi?ation" OR "aid agenc*" OR "international community" OR "civil society" OR "policy actor" OR "policy agent" OR "non-state actor" OR "global actor" OR "government" OR "international agenc*" OR "multi-lateral agenc*" OR "think tank" OR "stakeholder" OR "policy entrepreneur" OR "policy network" OR "policy maker" OR "decision maker" OR "practitioner" OR "state" OR "private sector" OR "corporate actor" OR "provider" OR "education business" OR "advocate" OR "coalition" OR "stakeholder" OR "education industry"

OR "PISA" OR "OECD" OR "ILSA" OR "international large-scale assessment" OR "Political economy" OR "government" OR "governance" OR "policy trajector*" OR "policy-mak*" OR "policy-shap*" OR "practitioner" OR "politics of education" OR "education polic*" OR "economics of education" OR "policy implementation" OR "policy borrowing" OR "policy lending" OR "agenda" OR "advocacy" OR "lobbying" OR "globali?ation" OR "globali?ing world" OR "westernization" OR "Europeanization" OR "world society" OR "global arena" OR "multi-level govern*" OR "multi-scalar govern*" OR "decision mak*" OR "policy-practise" OR "institutionalism" OR "critical theory" OR "neoclassic economy*" OR "neoliberalism" OR "neo-liberalism" OR "human capital" OR "knowledge economy" OR "knowledge market" OR "policy entrepreneur" OR "policy networks" OR "policy paradigm" OR "policy learning" OR "policy convergence" OR "policy transfer" OR "policy travelling" OR "global education marketplace" OR "capitalis*" OR "global network" OR "policy network" OR "network governance" OR "think tank" OR "policy communit*" OR "structures of power" OR "policy change" OR "education* reform")

AND TITLE-ABSTRACT-KEYWORDS ("school" OR "schooling" OR "education" OR "educational" OR "vocational training" OR "VET" OR "TVET" OR "professional training")