UNIVERSITÀ DEGLI STUDI DI PALERMO

Doctoral program in model based public planning, policy design and management

Department of Political Science and International Relationships

SSD SECS-P/07 – Business & Public Management

Using dynamic performance management to deal with behavioral distortions associated with school performance measurement systems

PhD candidate
ROBINSON STEVENS SALAZAR RUA

Supervisor & PhD coordinator
CARMINE BIANCHI

XXXII CYCLE - 2019

ABSTRACT

This doctoral thesis is aimed at analyzing behavioral distortions in public schools and outlining outcome-oriented performance measurement systems that prevent and mitigate them. An instrumental view of Dynamic Performance Management (DPM) is used to support that endeavor by 1) identifying how performance drivers impact on outcome and output end-results, 2) determining how end-results affect strategic resources, and 3) understanding how strategic resources and benchmarks define the dynamics of performance drivers. This approach is also used to foster a shift from an output to an outcome-based view in performance management, intending to pursue sustainable results in the long-term.

The case of the Colombian public schools has been analyzed to frame behavioral distortions associated with performance measurement systems. Test-based accountability systems are often used to assess school performance. An inconsistent design of these measurement systems may provoke the emergence of dysfunctional behaviors, which may jeopardize school outcomes such as educational quality. The case-study presented in this research illustrates how an excessive focus on improving test scores caused narrowing of the school curricula. In particular, a DPM simulation model has been built to show how an outcome-oriented view may support policy makers and system designers on dealing with behavioral distortions from the use of school assessment systems.

In order to design the simulation model, an experimental approach has been used in this research through the analysis of the following performance measures in the Colombian case-study. *Strategic resources*: Lower and Higher knowledge students; *Performance drivers:* Fraction of time allocated to traditional teaching, Fraction of time allocated to holistic education, Students knowledge ratio, and Balance and scope of learning ratio; *End-result:* Net change in Higher knowledge students.

The above variables are proposed, as an initial step, to assess Colombian public schools' performance through the lenses of a DPM approach. As a result, dysfunctional behaviors -that may emerge due to inconsistent test-based accountability systems- can be counteracted effectively.

TABLE OF CONTENTS

LIST OF FIGURES

1) INTRODUCTION

1.1) Problem relevance

Performance Measurement Systems (PMS) affect actions towards a direction that contributes to achieve relevant outcomes. Such systems are never neutral to individuals and organizations. An inconsistent design of PMS -regarding the system of rules in an organization (Borgonovi, 1996) and the outcomes of decision makers' actions- may generate dysfunctional side-effects (Smith, 1993; Bianchi and Williams, 2015).

In social contexts, there is a high risk that people adjust their actions to meet targets, regardless of the ability of such results to affect organizational outcomes positively (Bianchi and Salazar Rua, 2017). Benchmarking is a common practice to hold public organizations accountable for performance, however PMS can be manipulated to maximize individual pay-offs (Radnor, 2005; Smith, 1995). As Campbell (1969) stated *"The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures, and the more apt it will be to distort and corrupt the social pressures it is intended to monitor".*

Behavioral distortions occur due to inconsistencies in the design of PMS. These types of behaviors are often reinforced by rewards granted mechanistically if performance targets are achieved. It produces an increase in outputs and a decrease in outcomes in the short- and long-term respectively (Bianchi, 2016). In this context, dealing with behavioral distortions becomes crucial for the achievement of relevant outcomes. To this end, an approach aimed at fostering a shift from an output to an outcome-based view in performance management is required (Bianchi, 2012; Bianchi and Rivenbark, 2014). Dynamic Performance Management (DPM) is used in this research to meet this need.

This dissertation frames behavioral distortions, from the use of school performance measurement systems, through Dynamic Performance Management (DPM). The case-study of the Colombian public schools has been used to show how dysfunctional behaviors can be prevented and mitigated by using an outcome-oriented view in outlining school assessment systems. In the Colombian context, despite government efforts to improve educational quality, school performance has

been lower than the targets in national and international standardized tests. This has caused a growing concern among the educational authorities, who are looking for definite answers to this issue. Experts in the field assert that a low level in students' knowledge acquisition is related to unintended behavioral effects caused by test-based accountability systems. Such performance measurement systems influence teaching strategies in public schools for the purpose of attaining targets and rewards (Hamilton *et al.,* 2002). Hence, traditional teaching (i.e. type of education aimed at building basic notions in students through rote learning activities and exam preparations) is often prioritized over holistic education (i.e. type of education aimed at building complex learning skills through interdisciplinary projects, student portfolios and advisories). As a result, educational outcomes, such as educational quality, might be in danger because of goal displacement (Sills, 1957).

Concepts and experiences from the fields of performance management and education have also been integrated by using an instrumental view of Dynamic Performance Management (DPM). This view enables the identification of performance measures (e.g. strategic resources, performance drivers and end-results) to outline feedback mechanisms, which describe unintended behavioral effects. Likewise, efforts have been devoted to clarify how an inconsistent design of test-based accountability systems may induce schoolteachers to game the system. A DPM model has been used to explain the previous phenomenon and provide policy recommendations on how to enhance school assessment systems through an outcome-oriented perspective. These policy recommendations have been intended to counteract behavioral distortions that may emerge as a response to government pressure for increasing test scores and myopic compensation schemes. In addition, the Colombian case-study emphasizes the need for enhancing collaboration among distinct stakeholders in the education sector. This implies that shared strategic resources should be identified with the purpose of increasing the effectiveness of public policies. Thus, a consistent alignment between outputs and outcomes in the school context is crucial to improve educational quality.

The design of control systems (Flamholtz, 1983; Simons,1995; Ouchi 1979) has also been linked to the theory of street-level bureaucracies (Lipsky, 1980; Mintzberg, 1979) to set the conceptual roots on how to outline outcome-oriented performance measurement systems through DPM. On the one hand, from an

institutional perspective, this implies that DPM models should subtly embody the notions of both diagnostic and interactive control systems, the organizational structure, the organizational culture and the organizational environment. Likewise, DPM models should be designed by taking into account the characteristics of professional bureaucracies and clan controls. As a result, these simulation models could be used to foster feedback analysis, strategic dialogue, and information sharing in organizations. On the other hand, from an inter-institutional perspective, the attainment of outcomes requires that a multi-level and multi-actor approach is implemented. Such approach can be supported by means of an alignment between organizational outputs and community outcomes. Therefore, pursuing outcomes in the local area goes beyond the design and implementation of performance measures and policies in single institutions. DPM can be used as an outcome-oriented approach to support decision makers and system designers in outlining consistent performance measurement systems that foster a common shared view among distinct stakeholders.

In particular, DPM has been used in this thesis to support the design of school assessment systems in the Colombian context, intending to 1) deal with behavioral distortions caused by the use of inconsistent performance measures and 2) enhance the achievement of school outcomes such as educational quality.

1.2) Objective and Research questions

Objective

To use Dynamic Performance Management as an approach to analyze behavioral distortions in public schools, and to outline outcome-oriented performance measurement systems that may prevent and mitigate them.

Research questions

1) Can performance measurement systems trigger behavioral distortions?
2) What are the characteristics of such performance measurement systems that generate dysfunctional behaviors?

3) In which fields can behavioral distortions be found from the use of performance measurement systems? Why?

4) What is the contribution of Dynamic Performance Management to deal with behavioral distortions?

5) How can behavioral distortions be framed in the Colombian public schools?

1.3) Thesis outline

In order to analyze behavioral distortions in public schools and outline outcome-oriented performance measurement systems aimed at preventing and mitigating them through Dynamic Performance Management (DPM), this doctoral thesis has been structured as follows.

The second chapter comprises relevant literature on: Behavioral distortions from the use of performance measurement systems, School assessment systems, and the Colombian education system. Likewise, the scientific framework of this research is discussed. In the first instance, the conceptual roots of an outcome-oriented view in performance management are analyzed. To this end, the design of control systems in organizations is connected with the theory of street-level bureaucracy and clan controls. In the second instance, DPM is explained, and it is proposed as a consistent approach to support the attainment of sustainable outcomes.

The third chapter describes the research methodology. The rationale for adopting DPM, the research design, the model building approach, and the research outputs are examined. The aim of both second and third chapter is to provide the reader with the foundations for comprehending the analysis of the Colombian case-study, which is carried out in the next chapter.

The fourth chapter discusses the modeling strategy and simulation results. In the first instance, feedback mechanisms that describe the emergence of behavioral distortions -associated with school assessment systems- are outlined. Posteriorly, a DPM chart and a DPM model about the case of the Colombian public schools are analyzed. Both of them have been designed through an outcome-oriented view in performance management. The simulation results from the DPM model have been

used to propose sustainable policies aimed at counteracting behavioral distortions in the Colombian context.

The fifth chapter concludes this dissertation by 1) presenting a summary of the main discussions, 2) determining the contribution of the research to the existing knowledge, and 3) explaining the limitations of the study and providing indications for further research.

The appendices provide the reader with further resources to extend the analysis and discussions presented throughout this thesis. In Appendix A, the Colombian reward system is depicted. Such reward system mainly relies on test scores to allocate financial and social benefits to teachers and public schools. Appendices B and C portray the English version of two surveys that were designed to collect the primary data of this research. Appendix D reveals the equations that were used to build the DPM model. This information enables the reader to replicate the simulation results and increase transparency in the research.

2) LITERATURE REVIEW

2.1) Behavioral distortions from the use of performance measurement systems

2.1.1) Basic notions of performance measurement systems

Performance measurement is often perceived as mandatory for business success (Platts and Sobotka, 2010). This is the reason why different frameworks, methodologies and tools for performance measurement have been developed (Kaplan and Norton, 1996; Bourne *et al.,* 2002; De Waal, 2004). Using measurement systems aimed at measuring and tracking performance became popular in the private sector in 1980. Several scholars were pioneers in this movement by introducing the use of scientific methods and techniques into the businesses sector to support goal attainment (Kaplan and Norton, 1993; Deming, 1986; Forester, 1993). Such methods and techniques allowed system designers to identify quantifiable indicators related to organizational tasks, intending to make mission and vision statements explicit. In addition, these indicators enabled decision makers to determine performance over time and compare current values with benchmarks for the purpose of drawing conclusions. In the 1980s, this was an innovation to assess organizational performance, as the use of inspections was widely extended yet. The quantitative approach related to the measurements contrasted with the subjective and intuitive nature of the traditional inspections (McGinnes and Elandy, 2012).

Radnor (2005) defines performance measurement as gauging the results of an activity. Similarly, Simons (1995) conceives the performance measurement system as the formal routines and procedures used by managers to keep or change organizational patterns. Throughout the years, this practice -aimed at increasing efficiency and effectiveness- has gained a growing acceptation (Brigham and Fitzgerald, 2001) among distinct professionals (e.g. policy makers, consultants and academics).

De Lancer Julnes (2006) describes the components of any performance measurement system as follows.

- *Inputs:* These are the resources required to produce organizational outputs. In public education, classroom materials, teachers' experience and expenditures per student are examples of this measure.

- *Outputs:* They make up the final products or services that should lead to the achievement of the desired outcomes. In the education sector, scores in standardized tests are a popular output that is used to assess schools' performance and design public policies. In theory, high test scores should lead to an improvement in outcomes such as educational quality.

- *Outcomes:* These are the long-term consequences of the outputs in the whole system. They are often difficult to quantify due to their broad and complex nature. Examples of school outcomes are family involvement in the learning process and educational quality.

In addition, Ridgway (1956) developed a classification system based on single, multiple, and composite criteria to analyze the impact of using different measures on organizational performance. The criteria are explained as follows.

Single criteria imply that only one indicator is used to gauge performance. For instance, in the 1950s the Soviet industries measured organizational success as a function of monthly production targets (Berliner, 1956). Other performance measures were not significantly taken into account.

Multiple criteria denote the use of simultaneous variables to evaluate organizational progress. This type of criteria is based on the premise that organizational performance goes beyond the achievement of a single output. It looks for fostering the investment of efforts in a range of tasks that will allow the organization to grow balanced.

Composite criteria are used to calculate merged measures by assigning weights to each performance variable. For example, in Colombia, the Synthetic Education Quality Index (ISCE) is a merged measure used to assess public schools' performance. Such performance is calculated as an average result from four dimensions (i.e. progress, performance, efficiency and school environment), which have different weights (Zambrano, 2015).

2.1.2) Supporting organizations through performance measurement systems

In the literature, several frameworks aimed at supporting the assessment of organizational performance have been identified. Some of them include the balanced scorecard (Kaplan and Norton, 1992), the performance pyramid (Lynch and Cross 1991) and the performance prism (Neely *et al.,* 2001). Three decades ago, these frameworks represented an important innovation in the performance management field because performance was mainly associated with financial measures before their introduction. Including non-financial variables -such as those related to customers, internal process and innovation- represented a different view to understand the rationales behind a high-quality service delivery. However, societies all over the world have rapidly evolved, and in parallel, the complexity and magnitude of their problems have grown up. This has made that those priceless frameworks, which initially were an insightful answer to understand organizational performance, have lost their inner glow. One of the most important causes related to this phenomenon is their static nature (Todd, 2000), which contrasts with the dynamism of the current world. In these circumstances, novel research in performance measurement systems has arisen to support the decision-making process in modern organizations (Santos *et al.,* 2002; Youngblood, 2003; Li and Maani, 2011; Castellano *et al.,* 2004; Bianchi, 2016).

Performance indicators are prevalently used to assist decision-making in both public and private organizations. However, several scholars have claimed that such indicators are linear and ignore the effect of trade-offs in the system (Bianchi, 2016; Santos *et al.,* 2002). Others argue that performance measures induce individuals to be driven by a short-term view (Youngblood, 2003), which lacks a meaningful purpose for the entire organization. In the same line, further research suggests that indicators may 1) neglect the contribution of interactive control systems, 2) fail to support the process view, and 3) make people misconceive the role of benchmarking (Castellano *et al.,* 2004). As a result, unexpected over-reactions might be triggered and over-interventions might subsequently be implemented in organizational environments (Li and Maani, 2011). In this regard, Smith (1990) denotes how the extensive use of performance indicators to hold public institutions accountable may push managers to work through a narrow perspective. Under this position, measures

are often linked to political goals according to individual, rather than group, interests. The author also emphasizes that performance indicators may lead to a fragmental view on managing public services, which may make it difficult for decision makers to perceive the whole picture of the system. This may reduce the chances of dealing with complex social problems.

Decision makers should know the principle of opportunity-cost. In performance management, this implies that the choice to improve performance in a certain organizational area may cause a performance reduction in other areas, as multiple conflicting goals may exist. In the scientific literature, this concept is also known as “trade-offs” (Skinner,1969). Several scholars assert that trade-offs are inherent to organizational contexts (Santos *et al.,* 2002). Therefore, decision makers must comprehend how to select among diverse scenarios that might have positive or negative impacts on performance. Thus, a proper choice may lead to a consistent investment of resources, which may help organizations improve the delivery of outcomes.

2.1.3) Criticisms of using performance measures

Public organizations are expected to work in the delivery of community outcomes (e.g. quality of life, local attractiveness, educational quality, trust in government, civic-mindedness). The results of such organizations are made accountable by performance indicators. However, criticisms associated with the use of measurements have arisen as described below.

In the first instance, the relation between the use of performance measures and the successful performance in an organization is not entirely clear. Some studies have found positive effects of using performance measurement systems in organizations (Lingle and Schiemann, 1996; Gubman, 1998; Evans 2004). However, other studies have found partial or contrary effects (Neely *et al.,* 2004; Ittner, 2003; Austin and Gittel, 2002; Bourne and Franco, 2003; Smith, 1993). For instance, Neely *et al.* (2004) carried out a comparative analysis in two divisions of an energy distribution company. In one division, performance was assessed through the use of formal performance measurement systems. In the other division, flexible rules for performance evaluation were used. The authors of this study did not find real advantages in the level of sales and profits from using a particular approach to

gauge performance. Similarly, Bourne and Franco (2003) found that managers perceived activities related to performance measurement as negative to organizational performance because of its highly bureaucratic nature. Some managers even thought their employees should use time at work to develop new projects rather than spending it on data reporting. In education, the use of performance measures based on school inspections are controversial since positive and negative effects on the achievement of learning outcomes have been denoted in the literature (Klerks, 2013). On the one hand, several findings support the fact that using school inspections as an evaluation mechanism may positively influence the improvement of educational quality (Hanushek and Raymond, 2005; Luginbuhl *et al.,* 2009). On the other hand, some findings in the field do not relate inspections to an enhancement of school performance (Cullingford and Daniels, 1999; Shaw *et al.,* 2003; Rosenthal, 2004).

In the second instance, performance indicators neglect contextual factors that impact on organizational performance. Therefore, performance comparisons are often perceived as unfair because of external influences out of managerial control. For example, Van Thiel and Leeuw (2002) found that Dutch police officers were incorrectly judged to have low performance levels compared to past statistics. However, the scholars discovered that crime patterns in the Netherlands had also changed in the last decades, and such information had not been taken into account in the current measurement systems. In healthcare, Sipkoff (2007) found that using performance results to contrast hospitals from the British National Health Service (NHS) was not a wise decision, as demographical differences and typology of patients were not appropriately considered by the indicators. Similarly, in education, Brown (2005) stated that ranking public schools through performance statistics is arbitrary since these measures favor institutions in wealthy areas. In the same line, Bosker and Scheerens (2000) asserted that school league tables overlook students' socio-economic conditions. The previous two investigations denote that variations among school performance may reflect not only the use of distinct teaching strategies but also the existence of contextual factors that go beyond teachers and principals' responsibilities.

In the third instance, difficulties in selecting and agreeing in the most appropriate measures to be implemented in each organizational unit are recurrent,

as conflicting goals among agencies represent the norm rather than the exception. Consequently, to determine whether the organizational efforts are being addressed towards the achievement of community outcomes may be complicated (Adab *et al*., 2002; Hodgson *et al.,* 2007). This point of view is supported by research in the field that has detected a misalignment of performance measures used in different institutional hierarchies across public agencies (Micheli and Neely, 2010). As a result, such misalignment may lead to ambiguous goals which make it difficult to have a clear picture of how individual actions are linked to organizational performance (Buchanan 1974; Lan and Rainey, 1992; Perry and Porter, 1982; Rainey, 1993). Likewise, implementing reward mechanisms might contravene the improvement of the service delivery, as weak performance measures might be used (Han Chun and Rainey, 2006).

In the fourth instance, outcomes are difficult to be quantified. Therefore, they are often operationalized as a set of short-term outputs (e.g. solved crimes per year, the number of fines per month, students' scores in the standardized tests). Problems arise as politicians -in the attempt to improve results during their mandates- put excessive pressure over street-level bureaucrats to maximize outputs. It fosters the development of strategies that prioritize outputs over outcomes, which may be translated into dysfunctional behaviors in the long-term such as gaming and goal displacement (Bohte and Meier, 2000).

Last, dysfunctional effects -caused by performance measurement systems- have been detected in multiple domains (Goddard *et al.*, 2004; Pidd, 2005; Bird *et al.*, 2005). For instance, in mobility, the Shadow Strategic Rail Authority (SSRA) informed that British train operators missed stations to attain performance measures based on punctuality (SSRA, 2000). In safety, Her Majesty's Inspectorate of Constabulary (HMIC) stated that in England the number of crimes is even four times higher than the reported by the police (HMIC, 2000). This report 1) provided evidence of crimes that are not registered in the data-base until they are solved, and 2) brought to light intentional misclassifications of crimes in order to meet quotas. In healthcare, a study revealed that waiting lists are reduced in particular months by increasing the initial time to have an appointment with doctors (McCartney and Brown, 1999). Similarly, Werner and Asch (2005) found that some doctors prescribed cancer screening procedures to reach targets, even if such procedures

were not advisable. In education, experiments in schools illustrated how feedbacks may paradoxically cause detrimental effects on the achievement of long-term outcomes. Such effects have principally been observed during the execution of cognitive tasks (Visscher and Coe, 2003; Kluger and DeNisi, 1996).

2.1.4) Side-effects of performance measures

Performance measurement systems are based on the premise that holding individuals and organizations accountable leads to an improvement in the quality of services. However, this assumption is controversial because several side-effects have been identified. For instance, in the mid-1990s, Argyris (1952) conducted a study on behavioral distortions and its impact on the achievement of organizational outcomes. In this research, the effects of using budgets to measure organizational performance were analyzed through a case-study. Argyris reported that employees in a factory prioritized “easy orders” to meet quotas at the end of a financial period. This strategy triggered delays in the delivery of previous orders, as they were not finished in the same sequence in which they were received. This implied a violation of the procedures and rules in the organization (Jaworski and Young, 1992).

The above example -of dysfunctional behaviors associated with performance measurement systems- denote the importance of 1) identifying common side-effects caused by the adopted indicators in organizations (Smith, 1993), and 2) designing consistent performance measures to prevent and mitigate such detrimental effects (Bianchi, 2016). This may help policymakers understand the reasons that induce people to manipulate the system to maximize their pay-off, and consequently to formulate strategies that align individual and organizational interests.

Smith (1993) identified side-effects of using performance measures in organizations. These effects are described below.

- *Myopia:* A focus on short-term activities at the expense of long-term goals.
- *Ossification:* A reluctance to experiment new approaches.
- *Tunnel vision:* A preference for “easy to measure” targets.

- *Sub-optimization:* A fragmental approach that leads to an unbalanced growth of different organizational areas.
- *Measure fixation:* The focus of attention shifts from the intended outcome to the measured output.
- *Misrepresentation:* Fraud and making up results are used to produce an impression of outstanding performance.

In a similar vein to Smith, other scholars have also noted further behavioral distortions associated with the use of performance indicators. These are listed as follows.

- *Illegal acts and Falsification:* An intentional modification in the data to satisfy particular standards and requirements (Mars, 1982; Vaughn, 1983; Simon and Eitzen, 1986).
- *Biasing and Focusing:* Data reporting aimed at emphasizing favorable circumstances for the manager and the organization (Birnberg *et al*., 1983).
- *Smoothing:* A deliberate alteration of the data flow that does not affect the achievement of outputs in a given period (Ronen and Sadan, 1981).
- *Filtering:* Information withholding to avoid negative reactions from superiors (Read, 1962; O'Reilly and Roberts, 1974; Birnberg *et al*., 1983).

The above behavioral distortions have been categorized by Jaworski and Young (1992) as "Strategic information manipulations" because the performance measurement system is affected directly. This author has also defined a further category of dysfunctional practices called "Gaming performance indicators". This category implies that the processes are perversely affected but the rules are not broken. It leads to maximize personal benefits at the expense of organizational goals. "*Gaming of a performance measure is said to exist when the subordinate knowingly selects his activities so as to achieve a more favorable measure on the surrogate used by the superior for evaluation at the expense of selecting an alternative course of action that would result in a more desirable level of*

performance as far as the superior's true goal is concerned" (Birnberg *et al.,* 1983, p. 123, as cited in Soobaroyen, 2005).

2.1.5) Gaming performance measurement systems

Gaming is a dysfunctional behavior that has been defined in multiple ways. Smith (1995) describes gaming as an inclination to take advantage of the system of rules. This behavior results from using the wrong variables to assess performance. Similarly, Courty and Marschke (2007, p. 905) label gaming as "*undesirable responses that the designers of performance measures did not foresee*". This definition emerged as the result of empirical analyses carried out in federal job training programs. In this research, the scholars found that implementers developed program-compliance strategies aimed at improving performance outputs significantly (Courty and Marschke, 2004). In addition, Kelman and Friedman (2009, p.924) define gaming as a "*behavior that consumes real resources but produces no genuine performance improvement even on a measured dimension*". This definition implies the creation of a false image of organizational progress.

Gaming has also been conceived as a set of unintended managerial responses to accountability systems designed to support policy implementation (Terman and Yang, 2016). Such unintended responses are described in the literature as follows.

- *Threshold effect:* It implies that performance measurement systems encourage people to meet organizational goals but discourage them to exceed such goals (Bird *et al.,* 2005).

- *Ratchet effect*: This results from setting future targets through an increase of previous targets (Hood, 2006). Under this effect, people repeatedly report low performance levels to prevent managers from setting higher goals (Goddard *et al.,* 2004). This effect also makes people lack organizational commitment in the long-term, as they gradually look for reducing their investment of efforts.

- *Output distortion:* This dysfunctional behavior is denoted as "*hitting the target and missing the point*" (Hood, 2006, p. 516). This type of gaming implies that organizational outputs are achieved at the expense of community outcomes,

which compromises the real purpose of the measures. For instance, Oppel and Shear (2014) found that the United States Department of Veterans Affairs kept some veterans on a secret waiting list for months -before registering them into the patient tracking system- to meet the 14 days target.

In a similar vein, Bohte and Meier (2000) describe the concept of gaming as an attempt to manipulate performance measures. These system manipulations are detailed below.

- *Cutting corners:* It refers to a managerial approach that prioritizes quantity over quality to deliver outcomes. In this category, decision makers place their efforts in the production of outputs regardless of the impact of their actions on the whole system (Levine *et al.,* 1990). For instance, public schools may graduate students that have not acquired an adequate knowledge level to increase graduation rates. These dysfunctional behaviors often emerge in contexts where the achievement of outcomes is crucial but resource availability is constrained. As a result, organizations spread their efforts across a wide range of outputs rather than in a few of them, intending not to be perceived as unproductive (Bohte and Meier, 2000).

- *Biasing samples:* In this type of gaming, individuals and agencies only select and report cases with a high probability of producing positive results. This phenomenon is known as “cream- skimming” and it implies a perverse alteration of organizational efforts (Wilson, 1989). For instance, it has been found that some teachers focus their efforts on increasing the number of white people who pass university entrance exams while they reduce attention on minority pass rates, as the latter group usually scores lower in the tests (Bohte and Meier, 2000).

- *Lying:* This is a high-risk method of organizational gaming, as it implies a loss of trust between managerial and political divisions (Bohte and Meier, 2000). This dysfunctional strategy is often used to hide disadvantageous information and portray results that favor particular interests (Downs, 1967). In this regard, Meier (1985) claimed that the manager of *the Environmental Protection Agency’s Superfund toxic waste cleanup program* intentionally lied about her reports to show an impartial image of the inspections that were carried out in

toxic waste sites during the 1980s. Such reports were based on personal rather than objective criteria.

According to the scientific literature, the most common causes of gaming behaviors in public and private organizations are the design of poor performance measures (Greener, 2005) along with the use of output-oriented rewards (Heinrich, 2002; Courty and Marschke, 2004). These two causes may induce people to turn away tasks that require high investment of resources and devote more time and energy to those activities “easy to do”. For instance, Figlio and Rouse (2006) found that schoolteachers focused their efforts on the lowest performing students in the standardized tests, as their scores were easier to increase than the highest performing students. This strategy allowed them to receive performance bonuses.

On the other hand, the emergence of behavioral distortions may also imply a not-recognized problem. Gaming behaviors may also arise as a response to irrational goals set by politicians (Terman and Yang, 2016). In these circumstances, efforts to achieve short-term outputs are symbolic gestures to please a political agenda rather than to enhance the delivery of outcomes. However, this topic is beyond the scope of this research.

2.1.6) The relationship between gaming and output-oriented rewards

Heinrich and Marschke (2010) declare that distinct reward mechanisms aimed at aligning individual actions with organizational goals have been applied during the design of performance measurement systems. Social and financial benefits are linked to the attainment of outputs for the purpose of improving organizational performance. In particular, Rothstein (2008) carried out a detailed review on the use of compensation models in different organizational and geographical contexts. The scholar analyzed the use of performance indicators and rewards in: Soviet manufacturing enterprises, several transportation companies in Chile, the Britain’s National Health Service (NHS), and test-based accountability systems.

In parallel, psychological research has also been conducted on the use of incentives and its impact on building up motivation (Skinner, 1938; Deci *et al.,* 1999). In particular, the scientific literature distinguishes two types of motivation. *Intrinsic motivation*, which implies the enjoyment of an activity for its own sake. This type of

motivation is increased by using internal rewards (e.g. public recognition). Alternatively, people may have an *extrinsic motivation* that is enhanced by external rewards such as financial benefits (Platts and Sobotka, 2010). For instance, in healthcare, Davies *et al.* (2000) claimed that practitioners, from the British National Health Service (NHS), were primarily driven by internal rewards to improve the quality of the services. In education, several scholars argue that the use of incentives may boost performance in public schools (Hanushek *et al.,* 1994; Kemmerer and Windham, 1997; Odden and Kelley, 1997). Further research suggests that schoolteachers are mainly motivated by internal rewards and students' satisfaction (Kelley, 1999; Heneman, 1999). However, it does not imply a null preference for financial benefits. In fact, it has also been found that payment schemes related to performance may positively influence schoolteachers' behaviors (Ladd, 1999). The above studies show how the use of reward mechanisms has been a spread practice in multiple policy domains during the last years. Thus, rewards can be used to address individual behaviors towards the achievement of organizational and community outcomes.

On the other hand, side-effects from the use of rewards in organizations have also been evidenced. Kohn (1999) suggests the existence of a perverse effect of using rewards to assess performance, as people may get used to them. As a result, they may develop tactics to improve the "numbers". It provokes a loss of meaning in the adopted indicators. Similarly, Pollitt (2005) found that public sector employees learnt how to increase their scores in the measures for which they were made accountable. This scholar concluded that the power of indicators falls over time, and therefore they should be replaced periodically. Furthermore, Kohn (1999) also demonstrated how pressure -generated by incentive programs- may induce salespeople to game the system by prioritizing the sale of products for which they receive greater financial benefits. Likewise, the researcher proved that the chances of unethical behaviors are risen due to external incentives. Kohn's research outlines the dangers associated with performance measurement systems that strongly rely on rewards to address individual behaviors. These reward mechanisms may make people prefer "easy to reach" outputs at the expense of long-term outcomes.

In a similar vein, Platts and Sobotka (2010) argue that dysfunctional behaviors from the use of performance measurement systems are tightly related to reward

systems and staff motivation. According to research in the field, performance measurement systems that rely on external rewards may reduce intrinsic motivation in the long-term (Deci, 1971; Deci, 1972; Deci, 1976; Weiner, 1980; Deci, Betlly *et al.,* 1981; Levine and Broderick, 1983; Jordan, 1986; Shalley, Oldham *et al.,* 1987; Kreps, 1997; Deci, Koestner *et al.,* 2001). Therefore, the design of performance measures tied to external rewards may not be enough to achieve targets (Kunz and Pfaff, 2002).

In addition, gaming the system may emerge as a consequence of linking performance to both targets and budgets. It may eventually produce a counterproductive effect that may affect organizational performance (Jensen, 2003). Jensen argues that performance measurement systems -based on rewards and sanctions- make people inclined to lie and cheat, as punitive actions are forecasted if targets are not achieved. He also asserts that once people were forced to cheat due to the system of rules, they will rarely change their mindsets and they will even spread this dysfunctional behavior in further organizational dimensions. Other scholars argue that gaming behaviors may be conceived as a rational, rather than irrational, response to restrictive performance measurement systems that influence human relations. Exerting pressure -through control mechanisms based on rewards- may lead to an increase in organizational stress and tension. Moreover, if rewards are not distributed fairly, people will adjust their actions to favor non-cooperative settings (Ouchi, 1979). Thus, dysfunctional behaviors may appear from the adoption of performance measures that are incorrectly aligned to reward systems.

2.1.7) Behavioral distortions in public schools

The emergence of behavioral distortions in public schools -due to an inconsistent design of performance measures- is the main concern of this doctoral thesis. Hence, this section has been devoted to discuss previous research about dysfunctional behaviors in education.

Over the last decades, performance measurement systems have been used around the world to hold schools and teachers accountable for student achievement. In particular, the use of performance statistics has become an extended practice in public education (Goddard and Mannion, 2004). These statistics are reported regularly to the audience by means of league tables, intending to rank school

performance. This implies that public schools are classified according to their results in standardized tests (McGinnes and Elandy, 2012; Goldstein and Spiegelhalter, 1996). In particular, school decision makers and politicians often design and implement educational policies based on such tests and tables (Hamilton *et al.,* 2002).

School performance indicators are assumed to be self-evident. As a result, decisions that lack an exploratory view to improve school performance may be done (Strand, 1997). In addition, in the school context, measures recurrently embody a "carrot and stick" approach by which institutions are rewarded or sanctioned based on their results (Jacobs *et al.,* 2006). This approach is grounded on the idea that schoolteachers will improve service delivery if their efforts are linked to outputs.

Public schools contribute to the development of society by a high-quality education service. Schools are expected to promote cultural unity, enhance intellectual and economic life, and turn out "well-educated" citizens that support social and political initiatives (Tyack and Cuban 1995; Smith, 1998). Therefore, public schools are held accountable for a range of outputs such as graduation rates, test scores, and pass rates on the college entrance exams (Smith, 1994). However, the existence of multiple goals and performance targets raises the probability of an inversion between means and ends. This phenomenon is reinforced by limited financial conditions and high-pressure levels from external entities such as the government (Bohte and Meier, 2000). Likewise, perception of unfairness regarding how the measurement systems determine the organizational performance and how compensation schemes distribute social and financial benefits may rationally induce people to engage in behaviors that maximize their pay-off to the detriment of the service delivery (McGinnes and Elandy, 2012). Particularly, motivation to game the system may be an expected consequence of designing inconsistent measures (Bianchi and Williams, 2015). On this subject, Berman (2002) found that people learnt to use performance indicators as a means to get political and financial advantages.

It has been observed that dysfunctional behaviors -such as gaming and goal displacement- emerge because of a disproportionate focus on improving scores in standardized tests (Bohte and Meier, 2000). These scholars asserted that low-

income public schools in the USA were prompt to game the system to produce comparable results to those schools with better financial conditions. Moreover, Bohte and Meier claimed that incentive structures -aligned with performance measures- may lead people to cheat the system as an attempt to maximize their benefits. In the same line, Lynn (1996) affirmed that public agencies may intentionally execute actions that will enhance output-measures to the detriment of desirable community outcomes. It occurs because of performance and reward criteria that make people lose their purpose for the development of society (Blau and Meyer, 1971; Downs, 1967). As a result, reward mechanisms divert attention away from achieving organizational goals and foster an individualistic culture. For instance, Garcia (1995) detected an “alarming relationship” between correct answers and erasures on students' reading tests in an American public school. This behavior was an unintended consequence of a school performance program that offered financial benefits to schoolteachers and principals based on score improvements in standardized tests. Therefore, the nature of the reward system may influence school responses.

In a similar vein, in the 1990s, an improvement in student performance was observed after implementing reward mechanisms (Elmore *et al.,* 1996). However, similar effects occurred in other schools, while a suspicious shift in school strategies was implemented. This “strategic adaptation” resulted from the desire of increasing bonus payments (Kelley, 1998). Analogously, further side-effects in public schools (e.g. increase in stress and pressure levels, cheating in standardized tests, gaming the system, narrowing the curricula) have also been framed by the literature (Clotfelter and Ladd, 1996; Kelley and Protsik, 1997; Koretz, 1996).

The organizational culture may also influence the emergence of gaming behaviors as observed by Vasquez Heilig and Darling-Hammond (2008). The scholars carried out interviews to principals and teachers from Texas public schools, and they found that dysfunctional strategies (e.g. holding back and suspending students) were performed to boost scores in standardized tests. Using such strategies was usually justified by means of a culture based on assessments, evaluations, and performance indicators. In addition, the scholars found that some public schools refused the enrollment of low-performing students as they could affect the accountabilities negatively.

2.1.8) Recommendations to deal with behavioral distortions

A significant increase in accountability has been observed in the public sector over the last decades. In parallel, a raising number of behavioral dysfunctions -caused by inconsistent performance measurement systems- has also been detected. In order to mitigate detrimental behaviors to the improvement of organizational performance, several recommendations should be considered.

Performance indicators should reflect both tangible and non-tangible resources that affect the policy implementation in an institution. Likewise, stakeholders' interests should be made explicit during the design of indicators by using a participatory approach preferably. Moreover, pressure levels associated with the accomplishment of tasks and the attainment of specific outputs should be controlled strategically (Van Thiel and Leeuw, 2002).

Other research has also emphasized the need to improve integration between performance measures and compensation schemes (Courty and Marschke, 2003). Similarly, several scholars have recommended aligning performance programs with symbolic recognitions rather than monetary benefits (Heinrich, 2007). It fosters the development of an intrinsic motivation without possible side-effects (Frey and Benz, 2005). Moreover, experts in the performance management field have suggested using "enabling controls" for designing performance measurement systems, as such controls may increase transparency and flexibility by a multi-player approach (Cuganesan *et al.,* 2014; Adler and Borys, 1996; Ahrens and Chapman, 2004). According to these experts, "enabling controls" will make people feel less constrained by measures and more prompt to improve performance. For instance, in 2008, a study found that performance measurement systems that involved the participation of multiple-stakeholders -during the stages of design and implementation- mitigated the emergence of behavioral distortions (Wouters and Wilderom, 2008).

In the same vein, the judgment of performance only based on indicators may not reflect the real contributions of people in the whole system, as inherent characteristics to social issues (e.g. unpredictability) may not be considered correctly (Bosker and Scheerens, 2000). Therefore, making comparisons among players in a single system appears to be an obsolescent approach. Perhaps, using dynamic

indicators -from a combined institutional and inter-institutional view- may be conceived as an insightful answer to improve service delivery and collaboration among stakeholders (Bianchi. 2016).

This thesis proposes the use of the Dynamic Performance Management (DPM) as an alternative approach to support the design of consistent performance measurement systems aimed at preventing and counteracting the emergence of behavioral distortions. In particular, this doctoral research has used the case of the Colombian public schools to illustrate the benefits of implementing an outcome-oriented view in performance management to deal with dysfunctional behaviors.

2.2) School assessment systems

2.2.1) Basic notions of school accountability

In many countries around the world, school accountability is considered a central component for the development of government reforms (Levitt *et al.,* 2008). Several scholars define school accountability as a system to hold schools and teachers accountable for improving educational quality (Ladd, 2007; Hopmann 2008). In the education sector, the adequate acquisition of knowledge and skills is often operationalized by measuring student performance in standardized tests. The measurement of this output is highly related to reward mechanisms that look for an improvement in the service delivery by fostering positive individual and institutional efforts (Booher-Jennings, 2007).

An important distinction between two types of school accountability models has been denoted in the literature (Adams and Kirst, 1999; Firestone, 2002; O'Day, 2002; Garmannslund *et al.,* 2008; Levitt *et al.,* 2008). On the one hand, *external accountability models* -also known as bureaucratic accountability- are used to provide information about 1) the quality and the efficiency in service delivery, 2) the compliance of formal regulations set by the government, and 3) the use of financial resources. The main characteristic of these models is their top-down approach by which education service is subjected to measurement and control by national, regional and local agencies. In this category, formal authorities often use standardized tests to track student progress in public schools, and they allocate financial incentives based on scores. On the other hand, *internal accountability*

models are grounded on the idea that the education service should not be ruled through hierarchical divisions. Instead, schoolteachers should be made accountable by how they develop their professional commitments according to the opinions of other colleagues. This implies the use of a bottom-up approach by which the teaching practice relies on the criteria of experienced teachers that have already worked in similar contexts. Therefore, internal accountability models connect compliance of formal regulations with professional standards and codes of conduct, which are evaluated by professional associations and peer review mechanisms. International organizations -such as the World Bank- support the implementation of internal accountability models in public schools since these models are seen as feasible answers to deal with the deterioration of the education service in underdeveloped economies (World Bank 2006; World Bank 2017).

Assessment and *Evaluation* are widely conceived as synonymous terms, however experts in the educational field emphasize their differences. School assessment implies the collection of statistical data for making decisions on how to enhance the achievement of student learning goals. On the other hand, school evaluation implies the examination of procedures, curricula and materials aimed at supporting the delivery of a high-quality service (Harlen, 2007). Hence, while *inspections* belong to the evaluation category, the *standardized tests* belong to the assessment category.

The inspections should be seen as a complement to the standardized tests. Both evaluation and assessment methods gauge the achievement of learning outcomes through quantifiable outputs. The main difference between them is the extent and character of the measurements. The standardized tests are often administered at a national and regional level, while school inspections are more localized. The former allows policy makers to have a “whole picture” of school performance in a broad context through the identification of large-scale patterns. The latter provides a “local picture” of how public schools deliver the education service by taking into account their contextual factors. Moreover, the standardized tests are grounded on objective and consistent elements, while the inspections have a subjective character. Therefore, inspection mechanisms may be disproved by teachers, however such mechanisms can put in context the adopted indicators

through comprehensive judgments that reflect the intrinsic nature of the system where the schools perform their activities (Altrichter and Kemethofer, 2015).

In the category of school assessments, the literature distinguishes between two types of tests. They are *norm-referenced* and *criterion-referenced.* The former is used to compare learning accomplishment levels among students, while the latter is used to contrast the achievement of student outcomes with educational standards set by the government (Popham, 2003). Both type of assessments can also be used to measure the effectiveness of internal school policies and teaching strategies through individual data that reflects the acquisition of knowledge and skills by the students over time (Vaishnav, 2005). A further distinction between high- and low-stake assessments is found in the literature. *High-stake assessments* imply that the standardized tests are used to foster behavioral and instructional changes to improve school performance. To this end, rewards and sanctions are allocated among schools, teachers and even students depending on the scores got in national and international tests (Ryan, 2004; Haertel and Herman, 2005). On the other hand, *low-stake assessments* lack compensation schemes aimed at addressing institutional efforts in a particular direction. They have an informative character and they are administered in a periodic base without affecting street-level bureaucrats directly (Klein et al., 2000; Carnoy *et al.,* 2003; Jacob, 2005).

Formative and summative assessments have also been discerned in the literature (EPPI, 2002; Harlen, 2007). *Formative assessments* are mainly used by schoolteachers to identify learning gaps and adjust teaching strategies. These assessments use a holistic approach to track student progress in particular topics that have been taught during classes. On the other hand, *summative assessments* involve the achievement of learning goals set by the government, which enables students to advance in the next stages of the education system (OECD, 2005). Standardized tests belong to this category. The main characteristic of such tests is its homogeneity in the design and implementation. This implies that conditions for administration, questions, scoring rules and interpretations must be objective and consistent to make reliable comparisons among schools and students (Zucker, 2004; Popham, 1991). Traditionally, the standardized tests have been used as a tool to track progress on academic achievement.

Cultural and social factors influence the choice of school accountability systems (Bracci, 2009; Darling-Hammond, 2004). This implies that policymakers may decide whether to 1) measure certain outputs and outcomes (e.g. family involvement, teacher-student ratio, the number of students per classroom, graduation and dropout rates), and 2) use particular assessments or evaluation methods (e.g. holistic activities, standardized tests, inspections) depending on the needs and goals of a community. For instance, the American and Colombian education system strongly rely on standardized tests and external controls to hold schools accountable for student performance, while the Finnish education system relies on formative assessments to enhance both pedagogical processes and student learning (Kupianinen *et al.,* 2009). Likewise, Finnish schools have discretionary power and independency to design and administer their own curricula and assessments -in accordance with national regulations- to enhance educational quality (Niemi *et al.,* 2016). This implies that teaching practices are performed in a supportive and free environment, which fosters the development of instructional methods aimed at satisfying special needs of school students. This endeavor is supported by an extensive use of student portfolios, self-evaluations and learning feedbacks (Finnish National Board of Education, 2004; Rinne *et al.,* 2002; Hendrickson, 2012).

In particular, this doctoral thesis has mainly been focused on analyzing behavioral distortions from the use of school assessment systems. In this regard, the design of any test-based accountability system must consider the following elements (Hamilton *et al.,* 2002).

- *Goals:* These are the final outcomes of education (e.g. educational quality, civic-mindedness, social awareness). They are usually delineated as statements that express individual and organizational desires through competency standards.

- *Measures:* These are performance indicators that quantify the goals (e.g. scores in the standardized tests, graduation and drop-out rates). Therefore, it is indispensable that measures are correctly aligned with standards.

- *Targets:* These are school benchmarks designed to compare the current state of the system with a desired level. They are usually set on an annual basis.

- *Incentives:* These are positive or negative consequences that result from the ability of public schools to attain targets.

2.2.2) History of test-based accountability systems

Over the last decades, test-based accountability systems have widely been used to hold schools and teachers accountable for improving educational quality. Regardless the geographical context, these systems often share similar principles such as the allocation of rewards and sanctions aimed at fostering individual and organizational changes. However, their design and implementation may substantially differ depending on intrinsic factors (e.g. political agenda, institutional norms, relationships among stakeholders). Therefore, student assessments may also reflect political and economic interests from a local community (Kellaghan *et al.,* 2009).

The measurement of educational quality has been influenced by the development of sophisticated testing systems that may produce valid and reliable results (McDonnell, 2005). According to Hamilton *et al*. (2002), test-based accountability systems are a set of methods designed to allocate benefits and punishments -to schools, teachers and students- based on test scores. Such tests are usually administered in regional, national and international domains. Moreover, in an organizational level, high-performing schools may receive financial benefits and public recognition, while low-performing schools may be intervened or closed. In an individual level, high-performing students may receive rewards such as scholarships, while low-performing students may be retained in a particular school grade. These exams are mandated by the government, and they are designed by external parties to the schools (e.g. private companies). These tests are also characterized by having a homogenous format in the questions, conditions for administration, and scoring rules. However, such homogeneity does not imply the use of a particular type of question (e.g. multiple choice, essay writing) or type of test (e.g. norm-referenced, criterion-referenced).

Using large-scale tests and incentives to improve school performance dates back to the mid-1800s (Tyack, 1974). However, these tests became popular since the 20th century (Haertel and Herman, 2005). In the USA, the state of New York was a pioneer in the implementation of large-scale testing in public schools. This state

used test-based accountability systems, before their popularity grew up (Allington and McGill-Franzen, 1992). One of the first standardized tests, that was administered to high school students in 1923, was the Stanford Achievement Test (SAT-10). The purpose of this exam was to assess the effectiveness of educational programs and compare schools (Resnick, 1982). In the mid-1920s and the 1930s, the use of these assessments increased rapidly because they were conceived as a cornerstone in the development of education reforms (Haney, 1981).

Posteriorly, between the 1940s and the late 1950s, the standardized tests were mainly used to judge school curricula and assess organizational performance (Goslin, 1963; Goslin *et al.,* 1965). Tests were not used yet to hold schools and teachers accountable for student achievement. Likewise, incentives were not allocated based on test scores. However, the function of these assessments started to change over the years.

In the 1960s, the National Assessment of Education Progress (NAEP) was introduced. This was a periodic test that measured academic achievement in a representative sample of American students. Parallelly, the Elementary and Secondary Education Act (ESEA) was released, and compensatory education programs were implemented. In this period, the standardized exams were used as a mechanism to test the effectiveness of government compensations allocated to students with socio-economic disadvantages. According to some scholars, this was a step ahead in the use of standardized tests to support school accountability (Airasian, 1987; Roeber, 1988).

In the 1970s, the “minimum-competency movement” emerged as a model to determine whether a student had acquired the basic skills to be promoted to the next educational stage (Jaeger, 1982). Therefore, both students and teachers were hold accountable for performance, and the notion of testing as a trigger of changes in teaching practices appeared (Popham *et al.,* 1985).

In the 1980s, the “education reform movement” replaced the “minimum-competency movement”. The former movement surged because of an extended concern about educational quality. According to NAEP reports, students were failing to acquire the most basic competencies. In addition, performance in American schools was lower than other countries, despite the high investments in financial

resources. As a result, the format of the tests changed (it was mostly based on multiple-choice questions and few writing components), and student results in the exams were linked to rewards and sanctions to foster educational progress (National Governors' Association, 1989).

In the 1990s, a second "education reform movement" came up to deal with the criticisms associated with the dysfunctional effects that previous accountability systems were causing. A "score inflation effect" was detected because school hours were mostly devoted to exam preparations. Likewise, it was observed that the curricula were narrowed to focus the teaching practice on the tested subjects. This provoked an improvement in test scores, but it was not linked to better outcomes in education. During this movement, the concept of benefits and punishments for high- and low-performing schools remained the same. However, the design of tests was intended to help students achieve learning outcomes, despite teaching was inexorably focused on improving scores. Therefore, the tests -that emerged during this period- were used to assess higher order thinking skills (i.e. complex skills that are required to approach novel situations, such as critical thinking and problem solving). In addition, standardized tests were complemented with student portfolios, essays and teamwork. In this period, performance standards were often modified to satisfy community needs, and consequently, assessments were adapted to support such changes. In later years, the main adjustments in test-based accountability systems were 1) the participation of students with special needs during the administration of standardized tests, and 2) an increase in the contents and performance standards that were assessed (Hamilton *et al.,* 2002).

In 2001, the act of congress *No Child Left Behind* (NCLB) was released in the USA. This provoked a substantial increase in school accountability. Every state was mandated to assess their students in reading, mathematics and sciences. Likewise, school performance was officially determined by the AYP (i.e. Adequate Yearly Progress), which was calculated as a compound measure of students' test results. In addition, rewards and sanctions were allocated based on scores. For instance, high-performing schools received significant investments for classroom materials and their teachers received performance bonuses. Conversely, low-performing schools could be closed or be subjected to organizational changes (Figlio, 2005). This act of congress was controversial in the academic community since several studies

supported the benefits associated with high-stake testing (Jacob, 2002; Figlio and Rouse, 2004; Deere and Strayer, 2001), and other studies emphasized the side-effects of measuring school performance through the standardized tests (Jacob, 2005; Cullen and Reback, 2006).

In 2015, NCLB was replaced by Every Student Succeeds Act (ESSA), whose main purpose was to develop a flexible method to support school accountability. Therefore, failures of the previous education reform were approached by substantial changes in the system of rules (Korte, 2015; Darrow, 2016).

Nowadays, test-based accountability systems are the main mechanisms to track progress in the achievement of learning outcomes by monitoring annual improvements in specific outputs. Likewise, an emphasis on aligning competency standards with tests has been observed in recent years (Mathis and Trujillo, 2016).

2.2.3) PISA: A cornerstone in the development of education policies

PISA (Programme of International Student Achievement) is one of the most popular standardized tests around the world. This exam was designed by the OECD (Organization for Economic Cooperation and Development) to measure students' abilities for solving real-life problems (OECD, 1999). While other standardized tests (e.g. IEA, PIRLS and TIMSS) are mainly used to assess student performance in common subjects in the school curricula of the participating countries (Breakspear, 2014), PISA is used to gauge academic achievement through an innovative approach.

PISA results are conceived as a cornerstone in the development of education policies since its first administration in 2000. PISA is a comparative assessment that is administered every three years to 15-years old students in over 60 countries (e.g. the USA, Colombia, Finland, Italy, Russia, China, Indonesia). Each PISA cycle has an extension of six years and three exams are performed during this period. The first cycle was from 2000 to 2006, and the second cycle was from 2009 to 2015. In addition, the exams of each cycle have a different focus on testing domains. For instance, in 2000, reading was the main subject of the assessment. Therefore, a greater number of questions were devoted to it. In 2003, the attention was

concentrated on mathematics, and in 2006, the focus was on science. The same pattern was repeated in the second PISA cycle.

PISA has multiple-choice and open-ended questions that must be completed in two hours. Contextual factors are also taken into account by means of a thirty-minutes questionnaire about learning habits and students' socio-economic conditions. A similar survey must be completed by school managers to provide further information on the learning environment, teaching practices and educational resources. This information allows policymakers to put into context the measures. Hence, competencies -that are relevant for the integral development of citizens- are assessed by gauging literacy in mathematics, sciences and reading in students from different geographical and social contexts. Posteriorly, results are published through league tables, which are often used to support education reforms. The education systems of the countries ranked in the first and last positions are regularly named as high-performing and low-performing school systems respectively (OECD, 2004; Schleicher, 2007).

Several researchers have claimed that school performance may be improved through test-based accountability systems (Carnoy, 2001; Figlio, 2005). However, other scholars have associated the use of standardized tests with important consequences over street-level bureaucrats and students (Hershberg, 2002; Linn, 2000; Noble and Smith, 1994; Smith and Fey, 2000). Likewise, the beneficial effects of school assessments have also been rejected by some schoolteachers who argue that the standardized tests are irrelevant to improve learning outcomes (Shohamy, 2001). As a result, these assessments have not been implemented in all geographical contexts. For instance, Finland is considered a role model in education. This country has repeatedly ranked in the first positions in international tests such as PISA. However, its main characteristic is an education system that does not depend on standardized tests and school rankings (Gavrielatos, 2009). Similarly, in Northern Ireland, performance indicators are not used to assess public schools because of the possible side-effects that such measures may provoke (Bird *et al.,* 2005; McGinnes and Elandy, 2012).

2.2.4) Criticisms of using school performance measurement systems

In the scientific literature, the use of test-based accountability systems and school inspections have been associated with the appearance of behavioral distortions in public schools (Jacob, 2005; Cullen and Reback, 2006; Figlio and Getzler, 2006; Chapman, 2001; Ehren, 2006). De Wolf and Janssens (2007) have categorized such behavioral distortions into three categories.

The first category is related to intended behaviors such as window dressing, misrepresentation, fraud and reshaping the test pool (De Wolf and Janssens, 2007; Jones *et al.,* 2017). *Window dressing* consists of planning atypical lessons to satisfy inspectors' criteria. Findings on the topic suggest that this behavior was widely spread in England during the late 1990s and the early 2000s (Brimblecombe *et al.,* 1996; Case *et al.,* 2000; Chapman, 2001; Fitz-Gibbon and Stephenson-Forster, 1999; Wilcox and Gray, 1996). For example, in 1999, it was documented that 81% of British principals accepted that inspectors did not attend to conventional classes, as pedagogical methods were deliberately modified (Fitz-Gibbon and Stephenson-Forster, 1999). This action was executed to improve the chances of a successful evaluation in the inspection process. *Misrepresentation* consists of attempting to produce a favorable school image -which is usually distant from reality- during the inspection. In the Netherlands, it has been found that some schools included out-doors activities during the classes, intending to fulfill the minimum number of hours per lesson (Ehren, 2006). *Fraud* implies that schools and teachers adopt behaviors out of the system of rules that has been set by an external institution. For instance, Jacob and Levitt (2003) detected alterations in test-paper answers in a small percentage of schools in Chicago. *Reshaping the test pool* is a deliberate misclassification of low-performing students in the standardized tests or an encouragement to push out students who can negatively affect school accountabilities. For example, Schiller and Muller (2000) found that an inconsistent alignment between standardized tests and incentives may increase the chances of pushing low-performing students out of the testing pool. In this regard, research in the field suggests that greater school investments in human capital -such as the development of professional skills in the staff- may reduce the inclination to reshape

the test pool, as teachers may feel more confident to satisfy students' needs (Rustique-Forrester, 2005).

The second category refers to the emergence of unintended behaviors because of both assessments and evaluations. It includes 1) the use of teaching methodologies based on conservative rather than new approaches, and 2) narrowing the curricula through teaching to the test or teaching to the inspection. These distortions may lead to isomorphism in the long-term (De Wolf and Janssens, 2007), as schools may be ruled by criteria and parameters that homogenize their teaching practices at the expense of innovation (i.e. ossification).

The third category covers the other side-effects that do not belong to the previous classifications. On this subject, Perryman (2007) developed a theoretical framework that outlines how the pressure exercised by the inspection process may lead to unforeseen consequences in schools due to an increase in principals and teachers' stress levels. This side-effect may deteriorate the achievement of learning outcomes in the long-term.

In particular, a "score inflation" effect in test-based accountability systems has been found in the literature (Hamilton *et al.,* 2002; Sturman, 2003; Klein *et al.,* 2000; Tymms, 2004). This dysfunctional effect is the result of excessive government pressure to improve test scores. Such phenomenon leads schools to adopt traditional (notion-based) teaching -rather than holistic education- because of its supposed suitability to improve performance in standardized tests. However, this approach usually implies allocating more time to exam preparations, narrowing the school curricula, and rote learning. On the other hand, holistic education or "project-based learning" supports the acquisition of higher order thinking skills (e.g. critical thinking, problem solving and teamwork aptitudes). Such skills allow students to link the concepts learnt by attending traditional classes, and to apply them in real-life (Harel and Papert, 1991; Strobel and van Barneveld, 2009; Walker and Leary, 2009).

Pedulla *et al.* (2003) observed that some schoolteachers devoted more teaching time to the tested contents, despite they know the limitations of notion-based learning. This was the consequence of high-pressure levels for improving performance in standardized tests. In addition, Tymms (2004) and Klein *et al.* (2000) detected an increase in test scores in British and American schools as a result of

investing a substantial portion of school hours to train students for the exams. Similarly, Sturman (2003) found evidence on teaching to the test in British schools. Furthermore, Wiggins and Tymms (2002) carried out a comparative analysis between elementary schools in England and Scotland. They observed a greater incidence of dysfunctional behaviors from the use of performance measurement systems in British schools. This outcome was associated with the use of league tables to rank schools in England, which differs from the Scottish context.

Test-based accountability systems have also made schools and teachers prone to game the system by excluding students from the exams (Vasquez Heilig and Darling-Hammond, 2008; Darling-Hammond, 1991; Smith, 1986; Allington and McGill-Franzen, 1992; Figlio and Getzer, 2002). It has led to a substantial reduction in learning opportunities for minority groups -such as Hispanic and African American students- whose test scores tend to be lower than white students' results. In particular, the scholars argue that score increases in a testing year may be associated with an increase in retention and drop-out rates of the previous year (Allington and McGill-Franzen, 1992; Wheelock, 2003; Holmes, 2006). Additional research supports this position by claiming that high-stake testing provides greater incentives for gaming attitudes, which may lead to a higher number of students being retained or even dropping the schools out (Nichols *et al.,* 2006; Clarke *et al.,* 2000; Lilliard and DeCicca, 2001; Wheelock, 2003; Roderick *et al.,* 1999). In the same line, findings on the topic reveal that students who are retained, are more likely to drop out compared to students who are promoted (Heubert and Hauser, 1999; Rumberger and Larson, 1998). Schools may encourage low-performing students to abandon their studies by enforcing rigid attendance policies, implementing repetitive teaching formats, and neglecting special learning needs. Thus, an increase in test scores may be observed from reshaping the test pool, which may provoke long-term consequences in minority groups whose educational rights may be ignored (Owens and Ranick, 1977; Vasquez Heilig and Darling-Hammond, 2008).

In parallel, some researchers have detected that compensation schemes based on merit may induce individualistic behaviors in schoolteachers (Weibel *et al.,* 2010; Ballou, 2001; Firestone and Pennell, 1993; Kohn, 1986; Pfeffer and Sutton, 2000), which may be detrimental for building professional networks and improving organizational performance in the long-term. Findings on the topic suggest the

existence of an inverse relationship between an increase in external incentives and the inclination for sharing information and collaborating with others (Heyman and Ariely, 2004; Yang and Maxwell, 2011). Therefore, social and professional structures in schools may be deteriorated unintentionally by test-based accountability systems.

Similarly, research in the field suggests that informal networks may affect school performance (Siciliano, 2017). This implies that public schools with collaborative environments may display a higher level in teachers' performance compared to schools with less supportive environments (Jackson and Bruegmann, 2009; Pil and Leana, 2009). It represents a challenge for a fair allocation of external rewards from the use of test-based accountability systems, as equally talented workers may exhibit a lower performance due to contextual factors for which they should not be made accountable.

2.2.5) Recommendations to enhance school assessment systems

This doctoral thesis is mainly focused on the emergence of behavioral distortions associated with the use of school assessment systems. Therefore, this section portrays several recommendations - from the scientific literature- on how to enhance school performance measures and avoid their common side-effects. They are described as follows.

- *Aggregation of cohorts:* School performance is often assessed through test results in a given year. In the first instance, this action may provoke that outstanding schools -which have not been sanctioned in the past- may suffer punishments because of out of control circumstances. In the second instance, schools below performance targets may use dysfunctional strategies to game the system and achieve yearly outputs. Hence, aggregating cohorts over time may be an alternative to assess performance without punishing schools for annual results. It may also be conceived as a fair method to avoid sanctions in schools that are improving their performance consistently, but they are under the expected target yet (Figlio, 2005).

- *Aggregation of learning outcomes:* Figlio (2005) highlights that schools may be induced to narrow their curricula to only teach contents that are assessed in the standardized tests if additional subjects are added into the current

exams. In order to overcome this unintended effect, it is proposed merging distinct learning outcomes into a single indicator (Duncombe and Yinger, 1998). This action may reduce the attention on specific subjects and increase the use of a holistic education in schools.

- *Aggregation of student groups:* Several test-based accountability systems are designed to track student performance by taking into account ethnicity, socio-economic conditions and disabilities. This is done as a mechanism to promote equity in society, as the performance of minority groups can be followed. However, findings suggest that errors in the measurements are related to highly heterogenous schools (Kane and Staiger, 2002) because such schools may arbitrarily classify potential high-performing students as disables, intending to meet targets. Therefore, it is suggested to check whether a strict disaggregation of student groups is necessary in all cases in order to avoid this effect from the use of indicators (Figlio, 2002a).

- *Broadening the scope of measurement:* An adequate design of tests is central to enhance school accountability. Figlio (2005) remarks the need of aligning benchmarks with standardized tests properly, as it is expected that schools adapt their teaching strategies based on the exam structure (Figlio and Rouse, 2004; Jacob, 2002). As a result, schools may narrow the curriculum to teach specific contents and improve their outputs. Hence, they could receive rewards and avoid sanctions based on test scores. For instance, under the NCLB act, American schools that were sanctioned could lose a significant portion of their Federal Title I aid (Figlio, 2005). In order to cope with this side-effect, the design of standardized tests should consider broadening the scope of measurement by including a wide range of inter-disciplinary topics that foster the use of a holistic education. However, Figlio recognizes that an increase in the financial budget for education should be planned if this recommendation is taken into account. In the same vein, the scholar suggests that non-tested measures (e.g. graduation rates, drop-out rates, student mobility) are considered to calculate school rankings.

- *Using value-added systems for school assessment.* It consists of designing indicators that measure school performance based on improvements in an

extended time span rather than in a specific year. This system highlights the need of assessing the impact of teaching strategies on the achievement of learning outcomes. In addition, the use of value-added measures may reduce the incidence of dysfunctional behaviors, as pressure for meeting tight annual targets is released. However, interpreting results may become complex for the general audience because of the intrinsic need of tracking the educational strategies that were implemented during the previous years (Figlio, 2005).

In addition, some researchers assert that compensation schemes should consistently be adhered to test-based accountability systems (Stiefel *et al.,* 2005). This implies that both incentives and performance measures should reflect individual, institutional, and inter-institutional interests. In this regard, Siciliano (2007) argues that the effect of external rewards on internal competition and the influence of collaborative settings on individual performance should be conceived as central components in the design of school reward systems.

The emergence of dysfunctional behaviors in organizations is inherently associated with the intertwine between accountability systems and compensation schemes. Under certain circumstances, meeting annual targets may induce schools to behave improperly. However, it may partially be mitigated by rewards and sanctions that discourage such behavior. For instance, the American government compensates public schools by the foreseen number of disable students rather than the current number. This implies that misclassifying regular students as disables in order to meet targets in accountability systems may increase school expenses in the long-term, as it must be demonstrated that their special needs are effectively covered (Figlio, 2005). Thus, an independency between the standardized tests and the reward mechanisms in terms of their system of rules may contribute to a reduction of dysfunctional behaviors in public schools.

In parallel, a growing concern about enhancing higher order thinking skills (e.g. critical thinking, academic engagement and collaboration) in school students has been observed. In particular, it has been found a positive correlation between the use of school practices oriented to develop inter-personal and intra-personal abilities and an increase in graduation and college entrance rates, compared to traditional teaching formats (Rickles e*t al.*, 2019). Likewise, students who have

received a holistic education, have been associated with higher levels of competency in learning domains that require the application of basic notions in novel situations (National Research Council, 2012). In the same line, previous research in the field suggests that students who have been enrolled in schools that foster deeper learning opportunities -through inter-disciplinary projects, student portfolios, advisories, internships and technology integration- have improved their 1) grade point average (Collins *et al*., 2013), 2) test scores (Nichols-Barrer and Haimson, 2013), 3) college readiness (Friedlaender *et al.,* 2014), and 4) non-cognitive skills (Collins *et al.,* 2013; Guha *et al.,* 2014).

As a result, enhancing student achievement through a strategic alignment between traditional and holistic education may represent an alternative approach to build basic notions and higher order thinking skills to 1) overcome modern challenges, 2) succeed in civic life, and 3) reduce learning gaps among pupils with different socio-economic backgrounds (Rickles e*t al.*, 2019). Therefore, an outcome-oriented perspective in designing school assessment systems is crucial. This doctoral thesis proposes the use of Dynamic Performance Management (DPM) as an outcome-oriented approach to outline assessment systems in public schools.

2.3) Educational quality

The concept of educational quality is a matter of debate because of its complex nature. Some experts in education have declared that educational quality is the resource that allows children and youth to gain skills and abilities to overcome the modern challenges (Marquès Graells, 2001). Others assert that educational quality results from a pedagogical process by which organizational interests are aligned with student needs. Therefore, the education should 1) foster the development of students' skills, 2) provide schoolteachers with tools that support their teaching activities, and 3) encourage families to take part in the learning process of their children (Mestres i Salud, 2004). In the same vein, other scholars argue that the school education can only achieve a high-quality level if 1) it is designed to satisfy the demands and requirements of key stakeholders in a social system, and 2) the learning goals can be achieved satisfactorily (Escudero Muñoz, 2003). Escudero Muñoz also denotes that educational quality should be promoted by a sociocultural approach that maximizes the resources to deliver it, and guarantees a

wide access to education. Similarly, Schmelkes (1995) affirms that an ideal education system should provide students with an easy-access to educational resources. Likewise, it should be designed to enhance student achievement by fostering innovation and transformation in an individual and institutional level. Such transformation may be supported by the participation of families in school activities, which may stimulate an active learning environment.

The United Nations Educational, Scientific and Cultural Organization (2005) defines educational quality as a final outcome of school activities. Therefore, these activities should be designed and assessed carefully to enhance student learning. Likewise, the Organization for Economic Cooperation and Development (2010) describes educational quality as an *ex-post* concept to determine the potential contributions of students in their social system if they acquire an adequate level of higher order thinking skills. In the Colombian context, the Colombian Ministry of National Education (2013) has delineated educational quality through competency standards that allow policymakers to determine whether schools have achieved the minimum performance targets. Such competency standards are aimed at comparing the current state of the system with a desired goal. Hence, performance is measured across different school levels (i.e. pre-school, elementary, and secondary) in the subjects that students are expected to learn (i.e. language, mathematics, sciences and civic competences).

Competency standards are conceived as a guide to determine the type and quality of education that students may access in a given country. These standards support the adoption of instruments, techniques and methods to assess individual and organizational capabilities to achieve learning goals. This implies that competency standards represent the baseline to design curricula, school projects and teaching methodologies. In addition, competency standards set common criteria for both assessments and evaluations. For instance, standardized tests -that are aligned with competency standards- can be used to 1) monitor student progress over time, and 2) design teaching strategies focused on enhancing educational quality. Moreover, test results can be complemented with inspection processes to enhance performance by taking into account the socio-economic conditions where schools operate (Colombian Ministry of National Education, 2013; Cajiao, 2008).

The Synthetic Education Quality Index is used in Colombia to measure educational quality through four dimensions of performance (Zambrano, 2015). They are described below.

- *Progress:* It determines how much a school has improved its performance compared to the previous year. The main component to be tracked in this dimension is the change in the percentage of students who are in the lowest performance level in the standardized tests. The weight of this dimension in the final score is four points as a maximum.

- *Performance:* It discloses the average score in the subjects of mathematics and language that has been gotten by the students in the standardized tests. The higher the average test score, the higher the score in this dimension. The weight of this dimension in the final score is four points as a maximum.

- *Efficiency:* It denotes the proportion of students who are promoted to the next school level at the end of the year. The weight of this dimension in the final score is one point as a maximum.

- *School environment:* It is used to measure both classroom conditions and the relation between teacher and student. In order to quantify the variables of this dimension, a socio-economic questionnaire must be completed by the students after taking the standardized tests. The specific criteria to assess the school environment are not revealed by the Colombian Ministry of National Education. The weight of this dimension in the final score is one point as a maximum.

According to the Colombian Ministry of Education, this performance measurement system is designed to foster a balanced improvement in the four dimensions intended to gauge educational quality. Therefore, if a school devotes the organizational efforts to the enhancement of a specific dimension at the expense of the others, then its Synthetic Education Quality Index will not reflect an optimal result.

2.4) Description of the Colombian education system

The Colombian education system has a decentralized profile and its goal is to ensure the fundamental right of education to children and youth in the country. This system implies the existence of a sectoral organization where each government level has competences and complementary responsibilities to improve the service delivery. At the national level, the Colombian Ministry of Education develops policies and targets, establishes rules, regulates the education service, and tracks evaluations. It also provides public schools with technical and administrative help to strengthen their management capacity, and it allocates financial resources based on specific criteria. At the local level (i.e. departments and municipalities), the education secretaries manage the service by leading, planning, supervising and managing physical, human, and financial resources. They are also responsible for the results in terms of educational coverage and educational quality. Moreover, they provide public school with technical help, carry out teacher trainings, and apply rewards and sanctions (Codesocial, 2009).

The Colombian school system consists of the following academic levels.

- *Pre-school:* It is made up of three sub-levels. Pre-kindergarten, garten and transition. Children usually start this academic level when they are between three and six years old. The aim at this stage is to properly develop cognitive, social and emotional skills in children (Codesocial, 2009).

- *Elementary school:* It is made up of five grades. Children are usually between seven and eleven years old during this academic level. The aim at this stage is to build basic knowledge in languages, mathematics, sciences and civic competences (Codesocial, 2009).

- *Secondary school:* It is made up of six grades. Youth are usually between twelve and seventeen years old during this academic level. The main objectives at this stage are to 1) reinforce the notions that have been developed in the elementary school, 2) build higher order thinking skills to provide students with tools for overcoming novel challenges, and 3) prepare youth for their university careers or their inclusion in the market force (Codesocial, 2009).

According to the Colombian constitution, the education is a fundamental right for children and youth, and its effective delivery is a shared responsibility among the major social stakeholders (i.e. government, schools, families and enterprises). The Colombian constitution states that every student in the country must receive a high-quality education, regardless of his/her age, gender, race, religion, or economic condition. Public education in Colombia is mandatory and tax-free for children and youth between five and fifteen years old. In addition, The Colombian Ministry of Education allocates financial resources to the education secretaries based on annual percentages that are defined in political agendas. These resources must be used to 1) pay the salaries of schoolteachers and administrative staff, 2) finance the purchase of educational materials, and carry out maintenance and reparation activities, 3) support the development of school projects, and 4) provide low-income students with meals (Torres and Duque, 1994).

The four major laws that regulate the Colombian school system are described below.

- *Law 2277 of 1979***:** It outlines the payment schemes and the conditions for promoting teachers that have worked in public schools before 2002 (The constitution of Colombia, 1991).

- *Law 115 of 1994:* It describes the pedagogical guidelines that public schools must follow to deliver the education service effectively (The constitution of Colombia, 1991).

- *Law 715 of 2001:* In the first instance, it defines the tasks and responsibilities for which the national government and the local authorities are made accountable. In the second instance, it denotes the financial mechanisms that the government uses to support the service delivery (The constitution of Colombia, 1991).

- *Law 1278 of 2002:* It outlines the payment schemes and the conditions for promoting teachers that have worked in public schools after 2002 (The constitution of Colombia, 1991).

The Colombian Education System has four key actors: families, schools, government, and enterprises (Herrera Santana, 2007). An effective collaboration among such actors may generate behavioral changes in students in order to properly face the challenges in modern societies. Therefore, local agencies should perform their functions in a coordinated way to achieve important outcomes for society. Unfortunately, at present, the efforts of Colombian educational actors are not addressed towards the same path. First, the families are absent in the education process. Second, the link that should connect families, society, government, and business does not exist in practice (Llinás, 1995). These problems are caused by a fragmental and static approach in designing education policies in Colombia. Therefore, an outcome-oriented view in performance management is proposed in this research to outline such education policies through a holistic and dynamic perspective.

2.5) Standardized tests in Colombia: Background and type of assessments

In Colombia, state evaluations are designed by the Colombian Institute for the Promotion of Higher Education and monitored by the Ministry of Education. The standardized tests that are used to assess the Colombian schools are called “SABER” and “ICFES”. The former is aimed at verifying the progress in mathematics, language, sciences and civic competences of the students who are in the third, fifth, ninth grade of the education system. The latter is aimed at assessing a comprehensive knowledge acquisition by eleventh grade students before their entrance to the university system (ICFES, 2010). The format used in these tests is based on multiple-choice questions with single answers. Students must read a passage, interpret the information and select the most appropriate answer for the questions (ICFES, 2017a, 2017b). This format assumes that students carry out a logic process to reach the correct answer. In addition, the Colombian government has also administered international tests -such as PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study)- in public schools. These tests have been used to analyze the state of the education system by taking into account the socio-economic background where the Colombian schools operate (Cajiao, 2008).

The use of standardized tests is a recent practice in Colombia. Fernández Gómez (2005) states that ICFES has been administered in Colombia since 1968 in order to assess student achievement at the end of secondary school. Likewise, the scholar declares that SABER was first administered to a small group of Colombian schools in the 1990s. In order to ensure the reliability of the data collected by this standardized test, a representative sample was built. Schools from Bogota, Barranquilla, Cali and Medellin were chosen, as they were located in the most important cities of the country. This pilot test was successful. In the first instance, the participating schools received feedback to improve their pedagogical methodologies. In the second instance, the government could get a whole picture of how schools were performing their tasks to set the basis for future education reforms. Posteriorly, SABER started to be administrated in a large-scale, and it acquired a mandatory character for schools across the country since 2001. Language, mathematics, sciences and civic competences are the main knowledge areas to be assessed in this exam, as they are assumed to develop students' skills and aptitudes to overcome real-life problems. Nowadays, standardized tests are a cornerstone for the development of public policies in the Colombian education sector (Fernández Gómez, 2005).

Colombia's participation in TIMSS and PISA has strengthened institutional efforts to enhance educational quality. A brief description of their administration in Colombian schools is provided below.

TIMSS was first administered to the Colombian schools in 1995. The main goal of this test is to determine the mastery level in mathematics and sciences that students from all over the world have gained over the years. In addition, this test assesses the level of congruence between the school curricula set by the government, and the contents that are really taught in the classrooms. This exam has successfully been administered to students who are in the third, fourth, seventh and eighth grades of the Colombian education system. Relevant information for policy making has resulted from the analysis of this test. For instance, it was revealed that only 60% of the Colombian public schools consistently taught the curricula that were set by the Ministry of Education. In addition, only one third of the contents were mastered properly by the students. TIMSS results have also emphasized the need to improve the education service promptly, as the Colombian

students have repeatedly had low performance levels compared to students from the other participating countries (Fernández Gómez, 2005).

PISA was first administered to the Colombian schools in 2006. The main goal of this test is to assess student performance in mathematics, reading, and sciences. At this time, over 70 countries around the world take part in this exam. Whereas TIMSS is focused on the assessment of the contents that are taught in the classrooms, PISA makes emphasis on the assessment of students' skills and abilities to solve novel problems (e.g. assertive communication, critical thinking and problem-solving). PISA results are published by means of league tables, which allow decision makers to make comparisons among the participating countries and design education policies intended to enhance educational quality. Historically, Colombian students have had insufficient performance in this test, which has raised multiple concerns on how to improve this issue. Therefore, several assessment methods and education reforms have been designed in Colombia to increase student achievement. In particular, competency standards have been introduced to address teaching efforts towards the attainment of both outputs and outcomes (Colombian Ministry of Education, 2013). This implies that outcome-oriented methodologies (e.g. Dynamic Performance Management) intended to outline school performance measurement systems may represent a central component to enhance educational quality.

In Colombia, two types of assessments are mostly used to track student progress. They are explained below.

In the first instance, *summative assessments* are administered to measure performance through a benchmarking process. These assessments are used to compare student achievement with pre-defined learning targets. Serie of exams are periodically applied to determine whether students are promoted to the next academic level based on their scores (Scriven, 1996). In Colombia, summative assessments are usually administered through both *internal* and *external approaches*. From an internal approach, schools are free to measure student progress through exams that are designed by a board of teachers. These exams use specific performance criteria -which have been set by the schools- to assess the competencies of their students. From an external approach, the Colombian Ministry

of Education and the Colombian Institute for the Promotion of Higher Education design standardized tests (i.e. SABER and ICFES) that are administered in a large-scale. This type of assessment provides politicians and principals with crucial information to outline policies and strategies intended to positively affect school efforts towards the delivery of a high-quality service (Cajiao, 2008).

In the second instance, *formative assessments* are used to help students in their learning process through assignments, projects and other school activities. Whereas summative assessments have a direct implication on students' promotion and retention rates, formative assessments only have a pedagogical impact because they are used as diagnosis tools to examine students' knowledge acquisition from an individual perspective (Camperos, 1984).

In particular, the Colombian case -analyzed in the fourth section of this thesis- illustrates the emergence of behavioral distortions from the use of performance measurement systems that rely on summative assessments to gauge school outcomes.

2.6) Opinions on the use of standardized tests

Fernández Gómez (2005) states that standardized tests have become a cornerstone in Colombian education reforms since the last years. Test-based accountability systems have been used to support the enhancement of the education service. In particular, compensation schemes based on test-scores have represented a key component in such accountability systems, as they have been used to foster individual and organizational efforts towards the attainment of school outcomes. This scholar is in favor of using standardized tests to measure and track student achievement. He argues that standardized tests allow principals and schoolteachers to 1) know the competencies that students have gained over the years, and 2) develop school programs to reinforce strengths and mitigate weaknesses in the learning process.

On the other hand, Fernández Gómez also recognizes that misinterpretations of test results and ignorance of test limitations may deteriorate school performance in the long-term. For instance, comparisons among public schools -which have opposite social and economic backgrounds- are a common problem of using league

tables improperly. Therefore, this researcher provides several recommendations to better use standardized tests. In the first instance, he declares that these exams should only compare schools that operate in similar contexts. In the second instance, he recommends that schools do not adjust the teaching contents to the standardized tests. Instead, he suggests that enhancing performance in the standardized tests be conceived as a mean, rather than an end, in the learning process. In the third instance, Fernández Gómez stresses the importance of improving the quality of formative assessments by holistic activities (e.g. interdisciplinary projects and advisories). In the fourth instance, he advocates for a higher collaboration between principals and schoolteachers to identify student weaknesses and perform counteracting actions. Last, the scholar draws the attention to the government to design effective policies for enhancing equity and educational quality. To this end, he suggests that 1) school performance is not judged by standardized tests, as the impact of students' socio-economic background in test scores is difficult to be measured, 2) decisions based on league tables are avoided because school rankings promote unfair competency, and 3) reward allocation does not mostly depend on test scores since negative consequences might be expected.

Some scholars have also emphasized how test-based accountability systems may provoke relevant side-effects, which have the potential to weaken educational quality. For instance, the use of league tables and compensation schemes based on test-scores may provoke irregular behaviors intended to reduce performance gaps and increase the allocation of individual rewards (Hamilton *et al.*, 2002; Bosker and Scheerens, 2000; Brown, 2005). In a similar vein, Popham (2001) asserted that an improper use of the standardized tests may trigger consequences for the service delivery, such as investing a disproportionate time in exam preparations at the expense of other school activities. As a result, a score inflation effect may be expected. Therefore, scores may not reflect a real improvement in learning outcomes. Moreover, in the Colombian education system, bonus payments, public recognition and prioritization in the allocation of financial resources are some rewards that teachers and schools receive based on test results (See appendix A). In this regard, Lin (2000) declares that test-based accountability systems have lost reliability because their main components (i.e. standardized tests and incentives) have been associated with detrimental effects.

Hamilton *et al.* (2002) describe the most common behavioral effects that may emerge in public schools because of government pressure to increase performance in standardized tests. In the first instance, *narrowing the school curricula* to only teach the contents that might be assessed in the exams. In the second instance, *increasing exam preparations* to raise the chances of better results. And in the third instance, *mimicking the format of the assessments* to boost student confidence during the administration of national and international tests.

The measurement of school performance through standardized tests may represent a danger to the education system because these assessments rarely take into account the environment where the students develop their learning process (Martínez Rizo, 2010). This scholar argues that tests do not reflect how results were achieved, as contextual factors and pedagogical methodologies are not gauged properly. He also declares that the extended use of standardized tests is strongly linked to political interests. The enhancement of community outcomes, such as educational quality, represents a priority in any government plan. Therefore, transparency in accountability and adequate management of the education service is crucial. Standardized tests are a mechanism to satisfy these social demands.

However, in the Colombian education system, the key actors usually have a poor assessment culture, which leads to oversize the scope of the tests (Martínez Rizo, 2010). These exams are used as an absolute method to determine educational quality. In particular, policy makers disproportionately rely on them to design education reforms. As a result, ineffective policies are implemented in the schools, as decisions are made by using an approach that mostly measures students' knowledge acquisition through exams. This implies a prioritization of summative over formative assessments, and therefore student achievement is narrowed to test scores. Instead, Martínez Rizo argues that standardized tests should be seen as a complement to the teaching strategies because learning goals should only be measured by competent teachers who have tracked student progress during a relevant period.

Perhaps, two of the major limitations to implement Martínez Rizo's position are, in the first instance, that teachers with experience in designing high-quality formative assessments are required. Therefore, investments in human capital to

enhance teachers' skills in this field should be a priority. In the second instance, the use of formative assessments for external comparisons would imply to design new strategies for quantification and data aggregation, as formative assessments are intended to only measure students' knowledge acquisition in the classroom.

To sum up, positive and negative effects can be experienced from the use of standardized tests. In order to reinforce the positive effects, tests should be seen as an opportunity to support student learning. Likewise, this type of assessment should be used to make an early diagnosis of students and schools in risk, which also requires a greater support from the Ministry of Education. The identification of strategic resources, performance drivers and end-results -through an instrumental view of DPM- is proposed in this thesis to shed light on how to enhance school performance measurement systems. This is done for the purpose of pursuing sustainable outcomes in the long-term such as an improvement in educational quality.

2.7) Design of control systems in organizations

The design of control systems should take into account the context where the organizations operate (Flamholtz, 1983). This implies that a pre-defined set of targets and managerial procedures may not guarantee a successful policy implementation. In order to address efforts towards the achievement of community outcomes, a wider perspective in designing control systems is required. Policy makers often emphasize the need for monitoring, evaluating and adjusting actions to meet targets. Likewise, they regularly conceive the allocation of incentives as a natural consequence of individual and organizational performance. In the scientific literature, this detailed group of procedures and routines is named as *core control system* (Flamholtz, 1983) or *diagnostic control system* (Simons, 1995).

According to Flamholtz (1983), the core control system is only one of four components that should be considered in the design of organizational control systems. Figure 1 provides a schematic representation of such components.

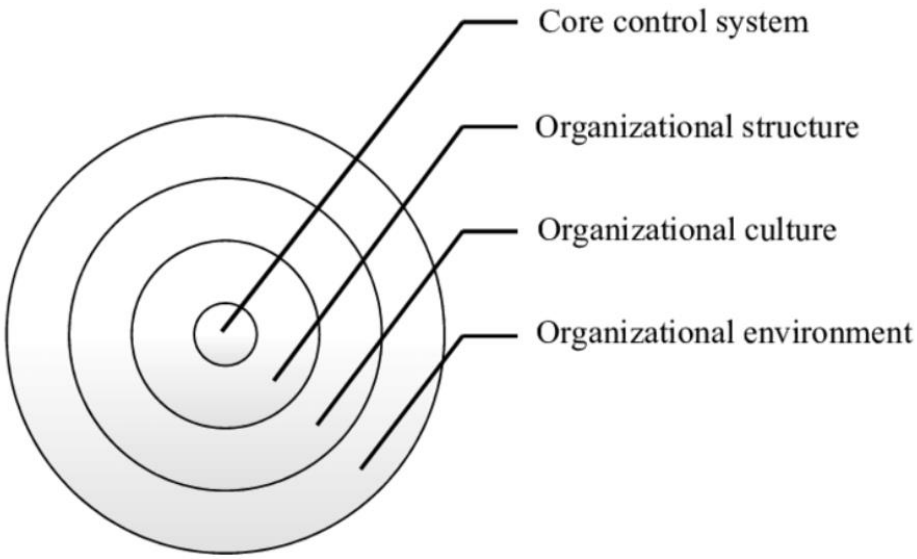


Figure 1: *Components of organizational control systems (Flamholtz, 1983; Maciejczyk, 2016)*

The core control system -represented by the innermost circle- is subdivided in the four components that are illustrated in Figure 2.

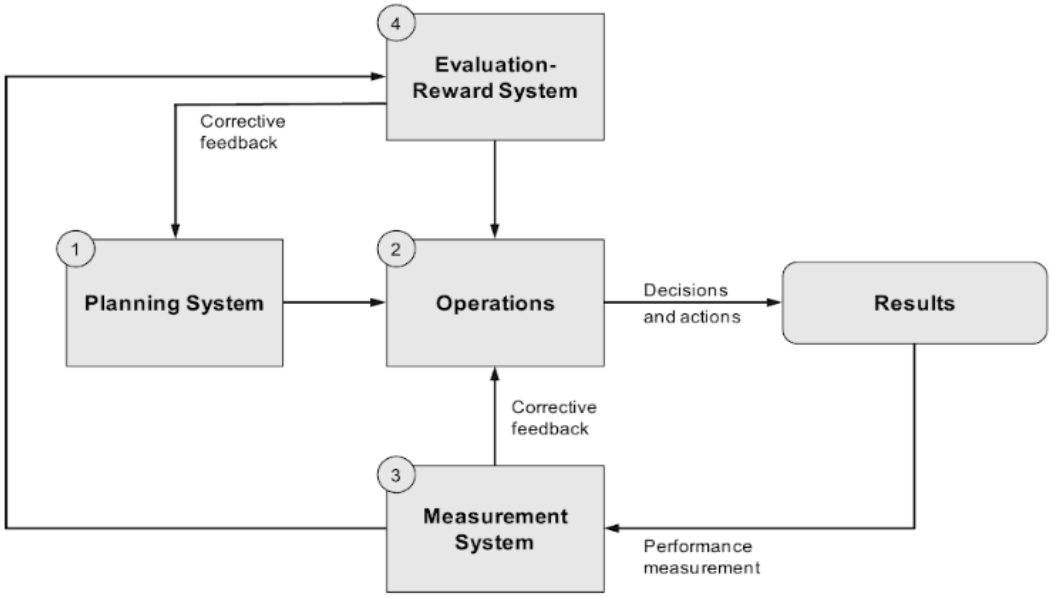


Figure 2: *Components of the core control system (Flamholtz, 1983; Kuhlmann 2012)*

The first component is *the planning system.* In this phase, organizational goals and strategic view are outlined to determine the steps to improve performance and achieve outputs and outcomes.

The second component are *the operations*. It refers to the tasks and activities that are crucial to achieve organizational goals.

The third component is *the measurement system*. In this phase, performance is gauged through operating results. Likewise, collected data is used to make short-term adjustments to the operations that are executed.

The fourth component is *the evaluation and reward system.* In this phase, incentives are allocated to distinct organizational units based on their results. In addition, long-term adjustments may also be carried out to redefine both operations and goals if necessary.

The design of control systems should also take into account the organizational structure, the organizational culture, and the organizational environment.

The organizational structure refers to the set of rules that allow the core control system to work properly (e.g. the use of a centralized or decentralized structure, and the workforce specialization by means of organizational units).

The organizational culture describes the values and beliefs that govern the organization (e.g. the institutional vision and the assumptions associated with the managerial processes).

The organizational environment is the space where the above components converge. It also includes the internal and external forces that may affect performance (e.g. individual behaviors, management styles, government reforms and stakeholders' pressure).

Several years after Flamholtz's research was published, Kaplan and Norton (1992) proposed the "Balanced scorecard" as a systematic framework to analyze critical performance factors in an organization. Such factors represented the key variables that should be monitored and assessed in any diagnostic control system, as they were the causes of either success or failure in policy implementation. Before the Balanced Scorecard was used in the public and private sector, performance was mainly conceived as the result of *financial measures* (e.g. operating income, return on equity, cash flow). Therefore, this framework represented an important change in organizational mindsets because the following *non-financial measures* were also

taken into account. *Customer measures* (e.g. customer satisfaction, market share, time delivery) were used to describe performance as a function of market conditions and customers' interests. *Internal process measures* (e.g. cycle time, unit-cost, capacity utilization rate) were used to enhance efficiency and effectiveness in organizational activities. *Innovation and learning measures* (e.g. training effectiveness, employee motivation, time to launch a new product) were used to encourage best practices and attitudes among employees.

Later, Simons (1995) remarked the need of complementing the diagnostic control systems -also named as core control systems- with the interactive control systems to pursue strategic transformations in public and private organizations.

In the first instance, Simons was in favor of using the Balanced Scorecard as a framework to support the diagnostic control systems by gauging critical performance factors. In particular, the measurement of such factors should 1) be verifiable, 2) capture relevant behaviors in an organization, and 3) reflect individual efforts (Lawler and Rhode, 1976). Unfortunately, control systems rarely meet such ideal conditions, and dysfunctional effects tend to emerge as a consequence of performance gaps (Simons, 1995). Measuring the wrong variables and relying individual performance on external incentives (e.g. financial benefits) are frequently the causes of a gap between the current and the desired state of the system. In order to reduce this gap, Simons highlighted the importance of conceiving the diagnostic control systems as a set of principles and processes that use negative -also called corrective- feedbacks for enhancing the service delivery. A classic example that illustrates how a negative feedback works is the temperature regulation in a home by means of a thermostat. If the current temperature is too high compared to a desired level, then the thermostat will attempt to reduce it.

In the second instance, Argyris (1977) and Simons (1995) suggested to shift from a single loop learning to a double loop learning. The use of diagnostic control systems -based on a single loop learning- implies that mental models, goals and performance targets are not questioned. Hence, the methods used in organizations to achieve performance targets are not altered. Once the organizational goals are set, pre-defined procedures are executed. It makes the organizations perform actions with a high level of autonomy and homogeneity, which eliminates the need of

communication among stakeholders. However, today's world is characterized by its fluctuating dynamics. It requires a high level of flexibility and innovation to approach novel events. In these circumstances, the use of interactive control systems -based on a double loop learning- that implies the participation of multiple stakeholders and fosters strategic dialogue is essential to review the underlying principles of organizational strategies. Therefore, interactive controls systems should also be implemented in organizations.

The interactive control systems embody additional features that complement to the diagnostic control systems. In uncertain environments, the design of adaptative strategies intended to align forecasts with organizational actions is a priority. In these circumstances, interactive control systems represent a powerful allied since a focus on processes -rather than pre-defined outputs and outcomes- becomes relevant to organizational performance. Likewise, these control systems stimulate dialogue and learning in organizations to enhance their operations. Moreover, the design of compensation schemes based on diagnostic control systems may be easier to manipulate than the design of rewards based on interactive control systems, as incentives in the former group are determined by fixed measures. Conversely, interactive control systems rely on subjective rewards that are allocated according to the superior's judgments about employees' performance. Therefore, innovative behaviors -which are difficult to quantify by using preset indicators- may be taken into account. It may also foster a positive -also called reinforcing- feedback based on information sharing among internal and external stakeholders (Simons, 1995).

In a similar vein, Ouchi (1979) argued that three types of control mechanisms can be identified and implemented in organizations depending on the clarity and congruence between their goals and the adopted measures. Likewise, the scholar remarked that no organization will display a pure control mechanism, as they usually embody particular characteristics of each type. Such controls are described below.

The market control is based on norms of reciprocity by which sanctions are executed if the parties do not behave properly during the financial transactions. In this control mechanism, people are rewarded by their precise contribution to the organization, which in turn implies that individual goals may be pursued if a reward

reduction is accepted. The market control is appropriate in situations where outputs can be identified and measured accurately, such as in economic activities influenced by price dynamics (Peterson's, 2010).

The bureaucratic control embodies not only norms of reciprocity, but also a legitimate authority that leads the subordinates' actions towards the achievement of organizational goals through a set of rules. Both performance assessment and agreement on hierarchical structures are major characteristics in this type of control. In addition, rewards are allocated depending on the superiors' interpretations about individual contributions to the organization. This is the most extended control mechanism in public and private organizations.

The clan control uses social agreements based on traditions, shared values and beliefs to foster knowledge exchange among people (Tighe, 2019), and to encourage individual efforts through limited supervision (Peterson's, 2010). This control mechanism is the most demanding to be implemented because 1) the employees' view should be congruent with the mission and vision of the organization beforehand, and 2) the managers should periodically reinforce the internalization of values and beliefs among subordinates by ritual and ceremonies (e.g. professional training). The clan control is usually used if performance control (Mintzberg, 1979) and behavior control (Govindarajan and Fisher, 1990) are not sufficient for the attainment of outcomes. In these circumstances, strict selection criteria and high levels of socialization are used to promote cooperative mechanisms intended to reduce goal incongruence between organization and individuals (Brenner, 2009). Thus, the clan control is appropriate in flexible settings where discretionary power is central, as different methods may be used to complete an activity. Likewise, it is suitable to support the achievement of community outcomes in *professional bureaucracies* (Mintzberg, 1979) or *street-level bureaucracies* (Lipsky, 1980).

Professional bureaucracies are usually found in contexts where the clan control is implemented. They do not depend on formal routines, regular supervisions from superiors, and other bureaucratic controls to work correctly. Instead, internalization and standardization of values and beliefs are crucial factors that ensure a proper service delivery to clients. Therefore, professionals have high discretionary power and work autonomously. They are also expected to achieve

similar results than in environments where constant audits are the norm (Mintzberg, 1979). In addition, professional bureaucracies often operate in a decentralized setting, which makes it difficult to plan a broad organizational strategy. As a result, performance is the product of individual initiatives that are characterized by an entrepreneurship spirit (Brock *et al.,*1999).

Street-level bureaucracies is a concept that was coined up by Lipsky (1980) to denote organizations whose employees have 1) discretionary power in the execution of their activities, 2) high level of autonomy from a legitimate authority, and 3) direct contact with citizens. In addition, these organizations often work in environments where the resources are limited, the demand of services is greater than the offer, and the goals are difficult to measure since they are ambiguous. Some examples of this type of organizations are schools, hospitals, police departments and tribunals, as their representative workers (i.e. teachers, doctors, policemen and judges) can decide the quality and the quantity of their contributions to the service delivery. This does not imply that street-level bureaucrats operate in a system that lacks rules. On the contrary, administrative officials and politicians articulate the standards and criteria by which performance is measured and behaviors are controlled. In this context, street-level bureaucrats are granted the possibility of pursuing the routes they consider most appropriate for the achievement of organizational outputs and outcomes. Thus, in a certain way, street-level bureaucrats may also be considered as policymakers since they are ultimately responsible for the process of policy implementation (Lipsky, 1980).

In street-level bureaucracies, the employees are usually less supervised than in other sorts of settings, as discretion requires freedom to be exercised. In addition, standards and indicators are defined in a context where goal ambiguousness is a key feature. The above attributes cause the results to be gauged without having clarity of the behaviors that were used to achieve them. This implies that quantitative measures rarely reflect the rationales by which people are judged because such measures do not support the qualitative content of the bureaucrats' actions (Broadkin, 2008). Therefore, a management-by-enabling, instead of a management-by-incentive, approach is required (Broadkin, 2011; Elmore, 1978; Lipsky, 1980).

In particular, several side-effects in street-level bureaucracies (e.g. goal conflicts and dysfunctional behaviors) emerge as a consequence of discretionary power and autonomy of employees, which hinders organizational controls. This condition would make one think that a removal of both attributes would lead to a final solution. However, the activities that street-level bureaucrats must perform in public and private organizations are complex, and therefore narrowing the job substantially or setting rigid rules are not an optimal response. Both approaches are not sustainable in the long-term because they cannot anticipate all behavioral scenarios that may appear during the execution of activities. For instance, in education, every child is different and has distinct learning needs. It makes the use of a detailed instructional approach almost impossible to be implemented in all circumstances (Lipsky, 1980). This implies that schoolteachers should use their discretionary power to decide the best teaching methodologies to apply during classes in order to help each student develop his/her skills and abilities.

The use of an outcome-oriented view in outlining school performance measurement systems through DPM is congruent with the ideas above discussed, and it allows to pursue sustainable outcomes in the long-term. In this research, the characteristics of the Colombian education system -such as a decentralized structure, clarity in the distribution of roles and bureaucratic procedures, high respect for education authorities, and some degree of flexibility to perform teaching strategies in the classroom- have been embodied in the DPM model implicitly.

2.8) Dynamic Performance Management (DPM): An instrumental view

DPM is used to frame problems with the use of output-oriented Performance Measurement Systems (PMS) in Colombian public schools, and to support the design of a robust set of performance measures based on an outcome-oriented view.

In particular, DPM can enhance PMS by trade-off analysis in time and space (Bianchi and Williams, 2015). Whereas trade-offs in time emphasize the effects of policies on different time horizons, trade-offs in space highlight the effects of policies on different subsystems. This distinction can help policy makers to 1) raise awareness about relevant delays in the system, 2) contextualize performance

measures, and 3) broaden system boundaries through an inter-institutional perspective (Bianchi and Rivenbark, 2014).

Figure 3 shows how DPM can support policy makers to outline sustainable policies by linking strategic resources with performance drivers, which in turn feedback into strategic resources through end-results. This approach can also improve the design of PMS by implementing modeling and simulation techniques.

DPM charts illustrate strategic resources as tangible and intangible stocks (Morecroft, 2007; Warren, 2008) that decision makers can use to develop their policies and meet performance targets. The availability of such resources varies depending on decision makers' actions. On the one hand, deliberate decisions (e.g. staff hiring) will affect the change of strategic resources that can be obtained from the market. On the other hand, the outcomes of the implemented policies (i.e. end-results) will affect the amount of such resources that cannot be acquired from the market (e.g. human capital, trust, image). These end-results are symbolized as flows in Figure 3. They accumulate in strategic resources, which are symbolized as stocks.

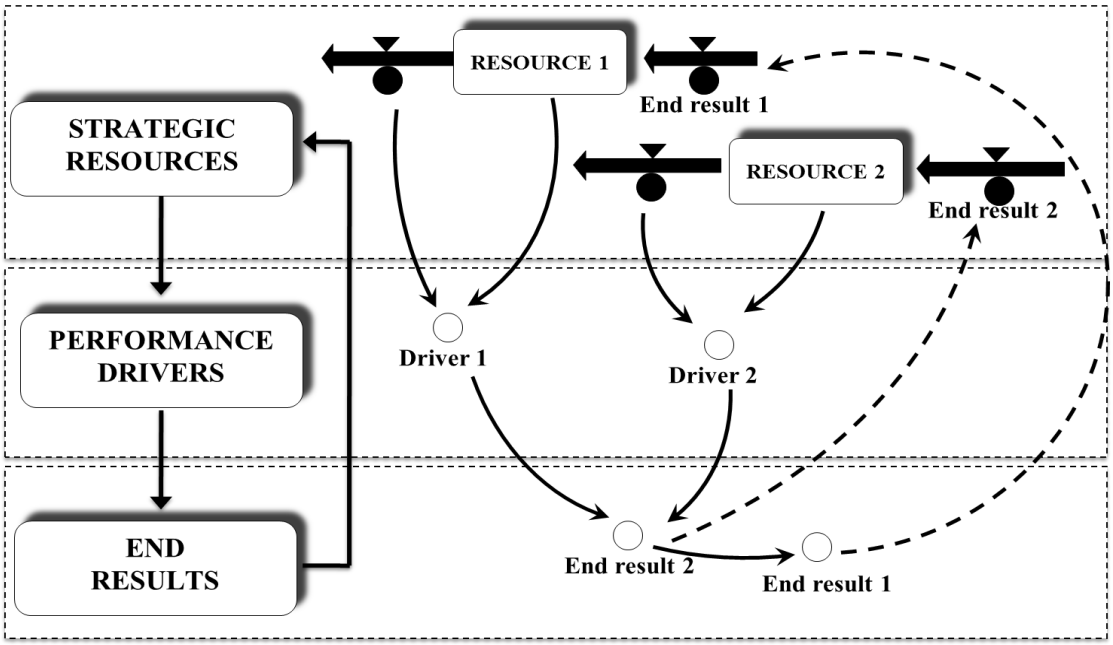


Figure 3: *An instrumental view of Dynamic Performance Management (Bianchi, 2010)*

To some extent, strategic resources can be controlled individually. However, if an unbalanced growth of them becomes the rule, then organizations will not reach their maximum potential. Changes in the strategic resources provide a partial view

about DPM. Therefore, critical success factors for achieving the end-results (i.e. performance drivers) must also be identified to understand how the system works from a holistic perspective. Performance drivers are gauged as ratios between the current and the desired state of strategic resources (e.g. the student-teacher ratio divided by a benchmark). These drivers should be measured and tracked continuously to recognize symptoms of failure and enhance the attainment of end-results (i.e. outputs and outcomes). Finally, end-results feed back into strategic resources and the cycle of performance is repeated (Bianchi et al., 2017).

Learning in-and-about complex social systems (Sterman, 1994) is possible through DPM. In particular, DPM describes organizational performance by making explicit the feedback structures, the mental models, and the decisions. This approach allows policymakers to deal with symptoms of crisis, such as the side-effects of an inconsistent design of indicators. For instance, schoolteachers may narrow school curricula by only focusing on tested subjects to improve student performance in high-stake tests. However, this strategy may be detrimental for the achievement of school and community outcomes in the long-term (e.g. educational quality, civic-mindedness, social awareness), as notion-based learning is prioritized over holistic education.

In the next section, the research methodology -which was used to analyze the case of the Colombian public schools and build the DPM model- is explained in detail.

3) RESEARCH METHODOLOGY

3.1) Rationale for adopting DPM

The DPM approach contributes to the enhancement of Performance Measurement Systems (PMS) by providing decision makers with effective tools to frame how and why indicators change over time for the implemented policies.

Traditional approaches in performance measurement often lead to a narrow and static view in data analysis and reporting, which rarely takes into account how targets are achieved. This implies that an output, rather than an outcome, view is mostly used to hold public organizations accountable. DPM can support policy makers in designing consistent PMS by using modeling and simulation techniques. In addition, DPM charts can be used as a framework to foster policy discussion by outlining a shared view among multiple stakeholders through the identification of strategic resources, performance drivers and end-results.

Furthermore, DPM can enhance the comprehension of how dysfunctional behaviors affect school performance through an outcome-oriented view (Bianchi and Williams, 2015). This view is crucial to avoid 1) an illusion of control as a result of a linear-thinking (Langer and Roth, 1975), 2) a goal displacement because of an inversion between means and ends (De Lancer Julnes, 2006; Sills, 1957), and 3) a performance adaptation since people learn how to game the measurement system, which causes indicators to lose their capability to differentiate between low- and high-performers adequately (Van Thiel and Leeuw, 2002; Meyer and Gupta, 1994).

In this research, DPM has been used to support the design of outcome-oriented performance measurement systems intended to prevent and counteract the emergence of behavioral distortions in Colombian public schools. To this end, a DPM simulation model -which takes into account trade-offs, non-linearities and time delays- has been built.

In particular, in the case-study of the Colombian public schools, a broader perspective in school performance measurement systems is needed to 1) extend the boundaries of performance indicators from test-scores to an assessment of holistic learning outcomes, and 2) align external targets -set by the government- with

measures that take into account the socio-economic background where public schools operate. An instrumental view of DPM is proposed to implement this perspective and pursue the achievement of sustainable outcomes. Thus, performance may be affected positively by focusing on end-results (i.e. outputs and outcomes), performance drivers and strategic resources.

3.2) Research design

A sequential exploratory design was used in this research. This is a type of mixed methods strategy that consists of two phases to gather the primary data. In this investigation, the first phase involved the collection of qualitative data by means of semi-structured interviews and a survey based on open-ended questions mainly. The second phase involved the collection of quantitative data by further interviews, a second survey based on numerical reasoning, and model validation sessions.

In addition, the secondary data was collected and interpreted by document review. In the first instance, open government data, local newspapers, and school performance reports were examined. In the second instance, analysis of the scientific literature in the fields of performance management and education was carried out. As a result, several simulation parameters were estimated based on secondary data (Luna-Reyes and Andersen, 2003).

In the first stage of this research, a preliminary field study -conducted through semi-structured interviews and a survey- was carried out to determine how stakeholders in education impact on educational quality (Salazar Rua, 2016). To guarantee the participation in this study of the main stakeholders in the Colombian Education System (i.e. families, schools, government, and enterprises), a non-probability sampling was used.

The sampling procedure is described below.

1) Colombian public schools -located in municipalities with different degrees of economic development- were chosen to carry out this research. For the highest economic development category, schools from Bogota and Barranquilla were selected. For the middle economic development category, schools from Monteria and Pereira were chosen. For the lowest economic

development category, schools from Sabaneta and Sabanalarga were selected.

2) In the above municipalities, a public school that met the following criteria was chosen. They were: 1) a student-teacher ratio and test scores in the national average statistics, 2) a medium-sized school, and 3) an educational offer for students in all the academic levels (i.e. pre-school, primary school and secondary school). A total of six public schools were involved in this stage.

3) Schoolteachers were asked to fill out a survey based on open-ended questions (see Appendix B). In addition, several teachers agreed to be interviewed in order to share their experiences. A total of twenty schoolteachers completed the survey.

4) Other stakeholders in education were also involved in this research. In particular, the sample included one person responsible to analyze test results at the Colombian Institute for the Promotion of Higher Education, two representatives from a local enterprise in Barranquilla, and four parents whose children study in the participating public schools.

In the second stage of this research, a survey -based on open-ended questions and numerical reasoning- aimed at understanding the effects of standardized tests on the behavior of Colombian schoolteachers was administered (See Appendix C). In addition, facilitated sessions were performed in a given public school, intending to receive further help in the model validation phase. A total of three teachers attended the sessions, and twelve teachers filled out the survey. In particular, the analysis of the collected data contributed to set the initial values of different variables in the DPM model, such as the teaching time allocated to traditional teaching and to holistic education, and the time affecting the inflow and outflow of the stocks of higher and lower knowledge students. Thus, an experimental approach in building the simulation model was implemented.

This research design was chosen because it was appropriate to 1) answer the research questions, and 2) triangulate the data in order to enhance the validity of the findings (Saunders and Lewis, 2012). In particular, the information collected -through surveys, interviews, school reports, open government data, and model validation

sessions- was used to estimate a range of values that performance variables and strategic resources would realistically portray in a medium-sized school in the investigated region. However, this was an exploratory research, which implies that the results should not be generalized since it was used a small sample of individuals.

3.3) Model building approach

Based on a preliminary field study carried out through semi-structured interviews with schoolteachers, prior knowledge of the problem context was obtained (Salazar Rua, 2016). Colombian teachers perceive the assessment system as a potential disadvantage for schools in poor areas. In fact, the assessment system does not consider how the socio-economic conditions of the areas where schools operate affect performance. In order to cope with this issue, schools have been inclined to increase their scores by allocating more teaching time to a narrow range of subjects, at the expense of others. As an unintended consequence of this unbalanced time allocation, the outcomes have been dropping despite slight increases in the scores. The risks of behavioral distortions, caused by such systems, require the use of proper methods that help Government design “robust” school performance benchmarking frameworks.

To start a dialogue with schoolteachers about the factors causing the described phenomena, two versions of a DPM model -that point out the interplay between strategic resources, performance drivers, outputs and outcomes- were built. These versions of the model were focused on the unintended consequences from the use of inconsistent performance measurement systems, and on the policy recommendations to counteract dysfunctional behaviors.

To this end, a focus on an empirical analysis -based on both primary and secondary data- was used. An exhaustive literature review, twelve surveys, three interviews with Colombian schoolteachers, document analysis, and model validation sessions with the support of experienced teachers, were carried out.

The model building process followed three steps.

In a first step, a problematic behavior was outlined by taking into account the literature review. In parallel, a causal loop diagram (CLD) was sketched to frame the

feedback structure that describes the emergence of behavioral distortions in public schools due to the use of standardized tests.

In a second step, a DPM chart was developed to illustrate how the system structure -portrayed by the CLD- may support an understanding of the effects of myopic education policies on school performance. Two sessions of three hours each were carried out to discuss both the CLD and the DPM chart with a group of three experienced teachers in a Colombian medium-sized school (about 500 students). The sessions raised several suggestions for improvement of both documents. The teachers, who attended the two facilitated sessions, were selected based on three main criteria: 1) qualification and experience, 2) involvement in school performance assessment, and 3) the status of the school, which in this case was of a public institution in a poor area of Barranquilla-Colombia.

In a third step, model structure and behaviors were validated -through group and individual meetings- to capture divergent ideas as suggested by Andersen *et al.* (2012). This analysis helped consider new insights, which led to iterative modifications in the initial model structure, parameter initialization, and simulation time. The validation of the model structure also contributed to test the hypotheses embedded in the CLD. In particular, building reliability in the DPM model was a gradual process that emerged as a result of the modeling experience during the facilitated sessions. Moreover, a qualitative, semi-formal and non-technical process was used to assess the usefulness of the simulation model to foster policy discussion (Barlas, 1996; Bianchi 2016).

Modeling strategy and simulation results were portrayed by following “the three C’s” (Zhang and Shaw, 2012). Therefore, *completeness* in data reporting, *clarity* in the process to attain research results, and *credibility* in such results by using a solid scientific reasoning, were central to this investigation (See Appendix D).

The primary goal of the model was not to replicate a reference mode with actual data because such data was not available in detail due to the recent adoption of the test-based accountability system in the country. The model was rather intended to raise the Colombian teachers’ awareness about the possible unintended consequences of policies focused on narrowing curricula to improve standardized test performance.

3.4) Research outputs: Data

As a result of the data collection phase, the following research outputs were produced.

- In the first stage of the research, twenty surveys were completed. In the second stage of the research, twelve surveys were completed.

- Three interviews with schoolteachers were recorded. Such interviews were intended to discuss participants' points of views and complement the information collected through the surveys.

- Over one hundred documents were collected and analyzed throughout this investigation.

In the next section, the case of the Colombian public schools is used to frame behavioral problems associated with the use of inconsistent performance measures through DPM.

4) MODELING STRATEGY AND SIMULATION RESULTS

4.1) Behavioral distortions associated with school assessment systems

According to the scientific literature, an excessive pressure to improve performance in standardized tests has caused the emergence of relevant behavioral distortions in public schools (Hamilton *et al.,* 2002). Such distortions are: 1) narrowing the school curricula to only focus the teaching practice on tested-subjects, 2) devoting a large portion of time in exam preparations, and 3) designing classroom assessments with a format similar to the standardized tests. These dysfunctional strategies may produce a "score inflation effect" -which implies a short-medium term increase in test scores but a long-term decrease in them- because higher-order thinking skills may not be built appropriately (Fuller, 2004 cited in Center on Education Policy, 2009; Linn, 1998; MassPartners for Public Schools, 2005).

In Figure 4, a causal loop diagram (CLD) outlines the effects associated with behavioral distortions in public schools because of perceived low performance ratios when static performance measurement systems are used.

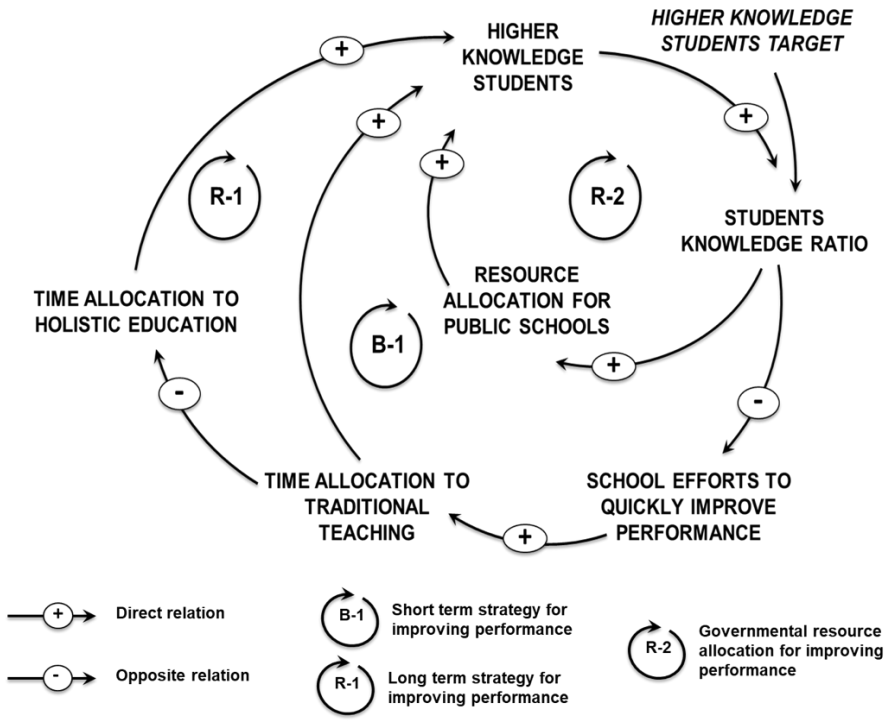


Figure 4: *Impact of government evaluation systems on school policies*

A perceived low performance ratio may affect how teaching time is allocated in public schools. “Students’ knowledge ratio” is the quotient between higher knowledge students (i.e. students who have acquired the knowledge and skills to approve the standardized tests) and a desired value. If such ratio decreases, schoolteachers must face high pressure levels to improve student performance. Therefore, in the short-term, an initial increase in the number of higher knowledge students (balancing loop “B1”) is perceived because of a “score inflation effect” (Hamilton *et al*., 2002).

However, the above results are not sustainable in the long-term since less time is allocated to pursue outcomes through a “holistic education”. Therefore, a reduction in the number of higher knowledge students will be observed, which in turn will lead to lower perceived performance ratios (reinforcing loop “R1”). The reinforcing loop “R2” shows how a decrease in the student knowledge ratio generates a further reduction in the resources that Government allocates to such schools. This generates a lack of investments, which would further cause lower performance levels in the long-term.

In brief, if school activities are primarily based on traditional (i.e. notion-based) teaching then a significant reduction in the time to build higher order thinking skills (e.g. critical thinking, problem-solving, teamwork aptitudes) will be perceived. As a result, the number of higher knowledge students will decrease in the long-term because the students will not be able to sustain high scores in the exams. Therefore, the acquisition of concepts in advanced stages of the student career will also be hindered due to a lack of accumulated knowledge.

4.2) Problems with narrowing the curricula in Colombian public schools

A DPM approach has been used to model how government evaluation systems -mainly focused on test scores and the assessment of traditional teaching outputs- led Colombian public schools to narrow the curricula. This situation may weaken student learning in the long-term despite of a perception of performance improvement in the short-term.

Figure 5 illustrates a DPM chart applied to the Colombian case. By following a bottom-up sequence, it identifies end-results to understand the impact of

behavioral distortions in public schools in the short- and long-term. It also identifies performance drivers influencing the accumulation and depletion processes of strategic resources. By mapping such resources, it supports an assessment of the gaps between the current state and the external targets set by the government.

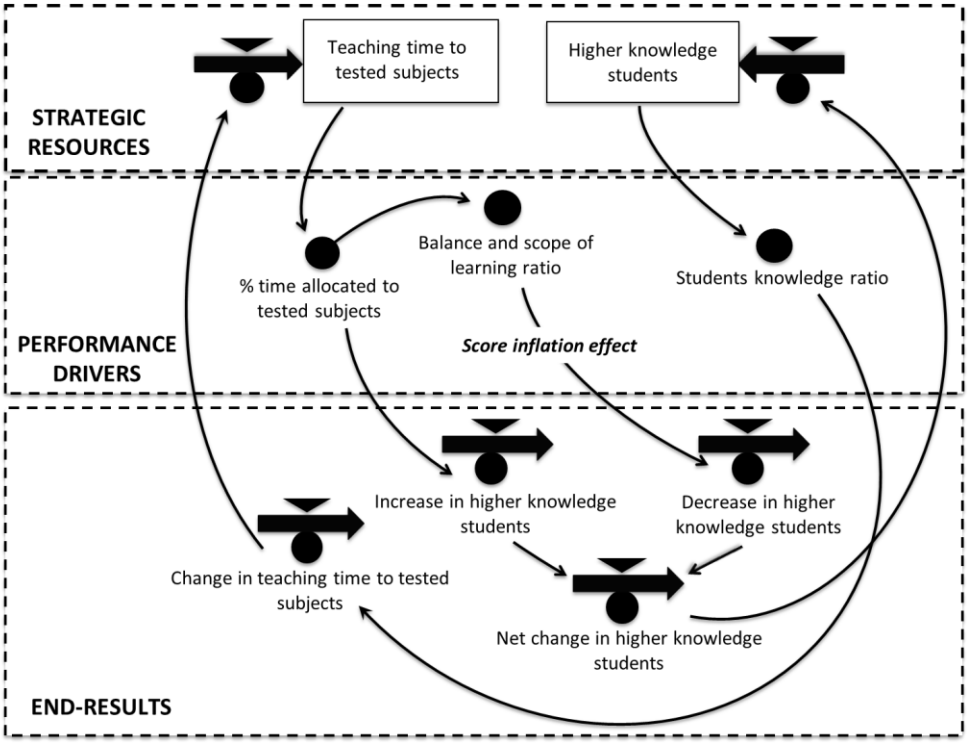


Figure 5: *Effects of myopic policies aimed at increasing the number of higher knowledge students in Colombian public schools*

A change in higher knowledge students is affected over time by the dynamics of the following performance drivers: "% Time allocated to tested subjects" (i.e. Number of hours allocated to tested subjects over Total school hours) and "Balance and scope of learning" (i.e. Number of hours allocated to tested subjects over Time allocation benchmark). Such drivers describe a policy through which a reduction in the student knowledge ratio (i.e. Number of higher knowledge students over a benchmark) leads to an increase in the time allocated to the tested subjects. When the student knowledge ratio is lower than one, schools react by looking for an immediate improvement in test scores. However, in the long-term, the side effects of this policy may negatively impact on the previous score gains. In fact, a higher fraction of time allocated to a bounded set of tested subjects (in proportion to the total teaching time) would increase and worsen the balance and scope of learning.

This would generate a “score inflation effect”, rather than an improvement in student aptitudes and skills.

In the long-term, also for those students who have initially been successful in their test scores, such “surface learning” (Marton and Saljö, 1976a, 1976b) could not sustain the achieved level in “higher knowledge students”.

Figure 6 shows a DPM simulation model designed to prevent and mitigate negative effects of school strategies based on an output-oriented approach. The main purpose of this model is to improve the understanding of complex dynamics rather than mimicking the behavior of historic time series (Bianchi and Winch, 2006). Therefore, the parameters it embodies and the results it generates through simulations should not be taken as accurate estimates. Instead, the model can be used to develop an understanding of how feedback system structures affect system behavior.

Figure 7 portrays the results from a simulation run of a DPM model based on the previously discussed myopic policies. The simulation has been run over a 120-months’ time horizon (10 years). This time extension has been set to capture the unintended side effects of such a policy on students learning outcomes.

The DPM model illustrates the case of a ‘generic’ medium-sized Colombian school in a poor area. This school has 50 students in a given cohort. It will take 10 more years for the students of such cohort to accomplish their studies. The purpose of the model is to portray the risks of poor outcomes related to teaching policies that aim at attaining significant improvements in standardized test scores in the short-term. Such risks may occur due to a bounded capability of students to sustain high test scores levels over the years, if only traditional teaching methods are adopted.

Despite the high level of synthesis used in this model, it was experienced a change in schoolteachers’ mindsets on how to attain sustainable goals from their efforts to improve standardized test scores. Discussing the model feedback structure and behavior with the participants by means of facilitated sessions, enabled them to perceive the need of identifying and monitoring the drivers of unintended outcomes from myopic education policies. Through the model, teachers could reflect on how silent phenomena -caused by the adoption of irrational policies- may generate poor

results in the medium-long term (e.g. prioritizing rote learning to the detriment of holistic education in order to increase scores in standardized tests). Therefore, keeping the time behavior of such performance drivers under constant control may help schoolteachers detect early symptoms of unintended and unsustainable outcomes of education policies in order to counteract them promptly.

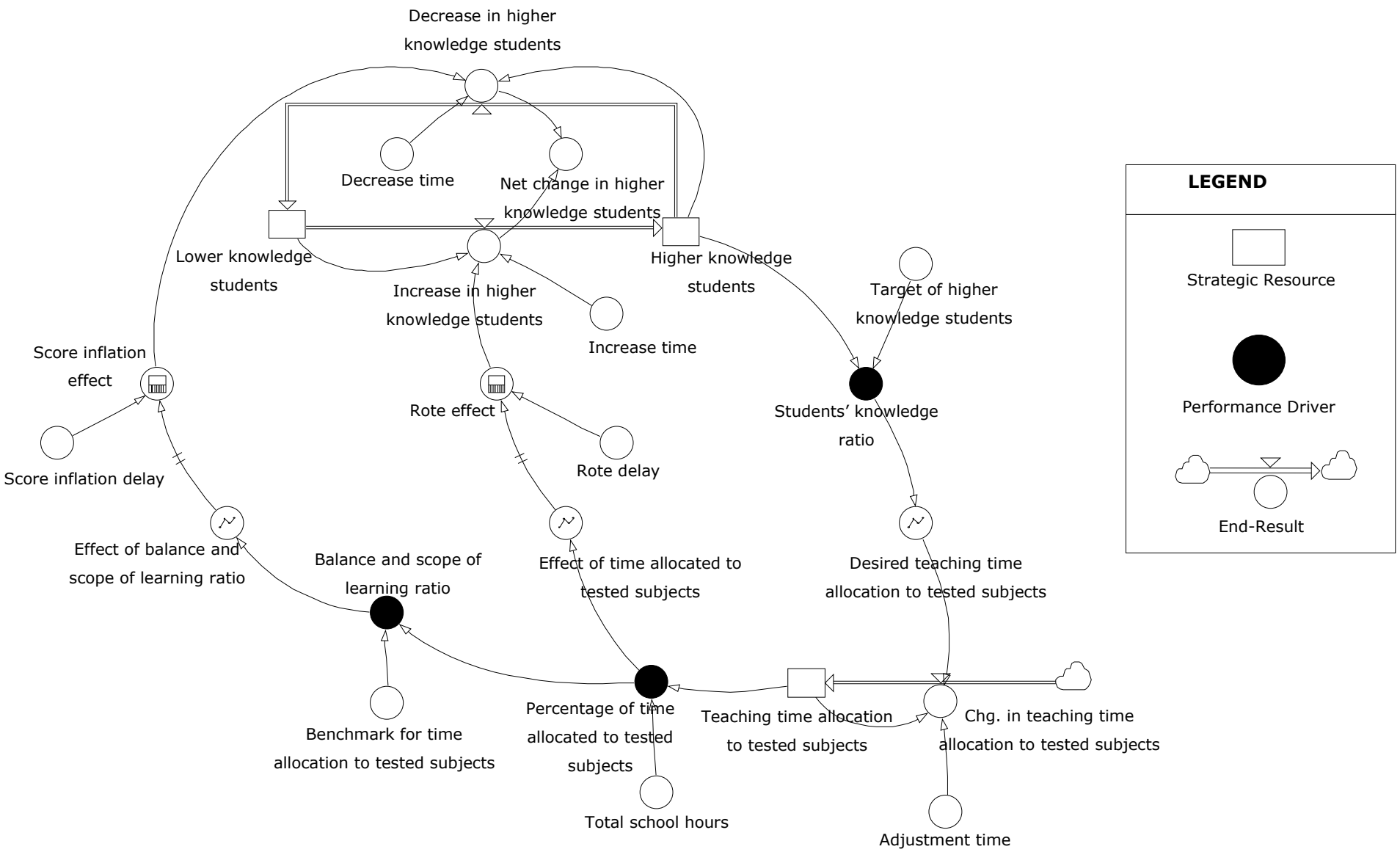


Figure 6: *A DPM simulation model illustrating the effects of myopic policies aimed at increasing the number of higher knowledge students in Colombian public schools*

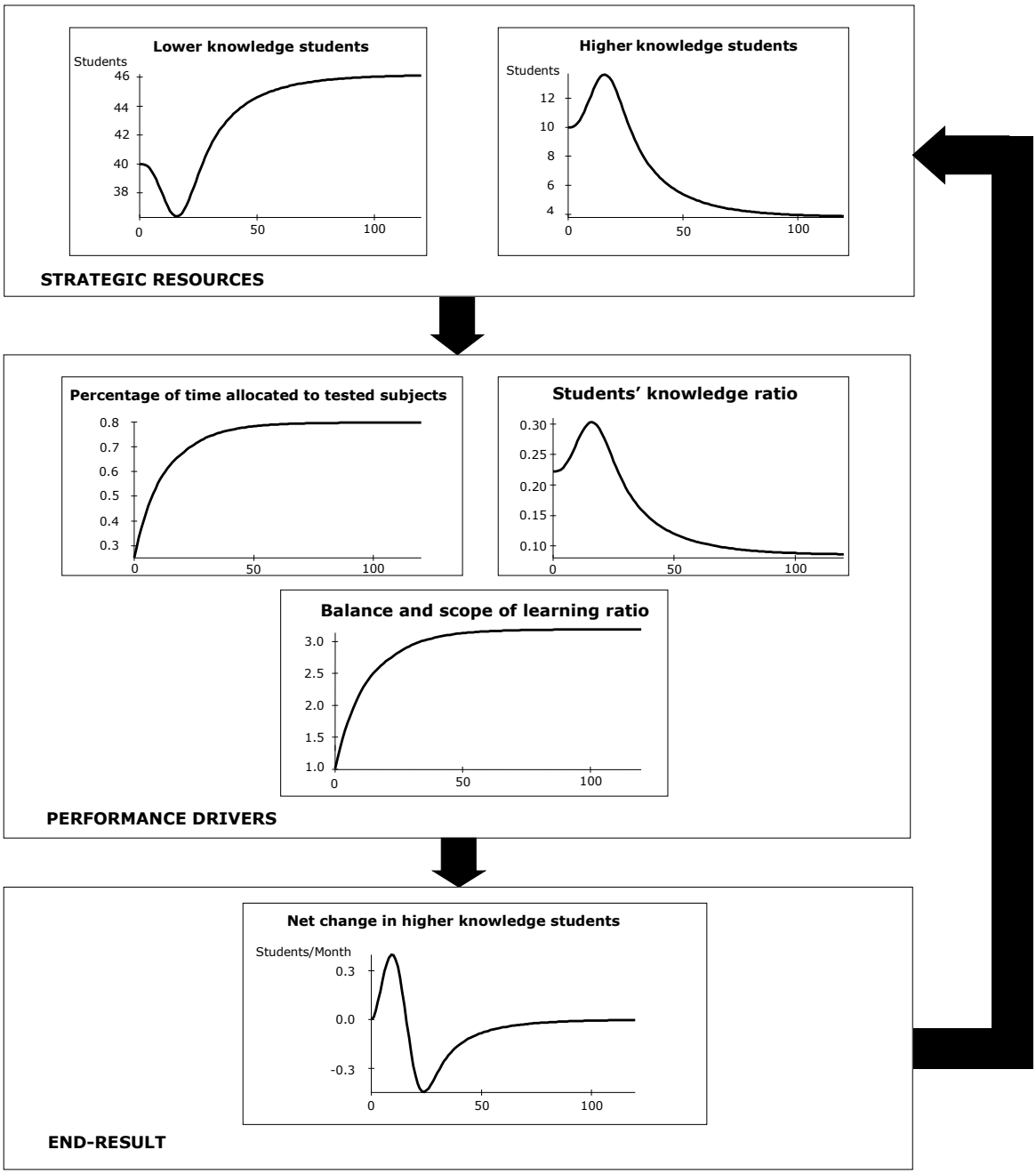


Figure 7: *Results from a simulation run illustrating the effects of myopic policies aimed at increasing the number of higher knowledge students in Colombian public schools*

The school described by the model has been unsuccessful in the past in achieving its targeted level of higher knowledge students: only 20% of total students passed the standardized tests. In order to improve performance and pursue a significant increase in the fraction of higher knowledge students over the total population, the school increases the time allocated to tested subjects. In the simulation model the target has been set to 90%, intending to illustrate how

Colombian schools in poor areas may emotionally respond to the gap in test results by setting overambitious goals that lead to poor outcomes in education.

As an effect of this policy, in the initial sixteen months the number of students who passed standardized tests increases, leading to a positive and rising net change in the number of higher knowledge students. It leads to an improvement in the "students' knowledge ratio". However, since the level of such a ratio is still below the target, the school continues to divert the teaching time to the subjects on which standardized tests are done. This implies a prioritization of traditional teaching over holistic education. This phenomenon is captured by the increase in the "Balance and scope of learning ratio": the higher this ratio is, the lower students' holistic learning will be, because of the diminishing residual teaching time allocated to inter-disciplinary projects, teamwork, and student portfolios. As an outcome of this side effect, after the ninth simulation month, the net change in higher knowledge students progressively decreases, and becomes negative after the fifteenth month. Such unintended outcomes are only latent until the sixteenth month. After that time, they become clear through a progressive decline in the number of higher knowledge students.

The difficulty to perceive such phenomenon is associated with the feedback structure of the system, and namely with the delays affecting the two flows that lead to the net change in higher knowledge students. It is also associated with the diminishing returns of additional teaching hours allocated to tested subjects. The understanding of such counterintuitive behavior portrayed by the dynamic complexity of the analyzed system is a benefit of using a DPM model to enhance performance management.

4.3) Policy recommendations

This section has been devoted to outline a broader set of performance measures and targets intended to prevent and mitigate the narrowing of curricula in Colombian public schools.

In Figure 8, the model in Figure 6 has been extended to show how an outcome-oriented view in performance management may positively affect the teaching time allocation in public schools. It has been proposed to assess outputs

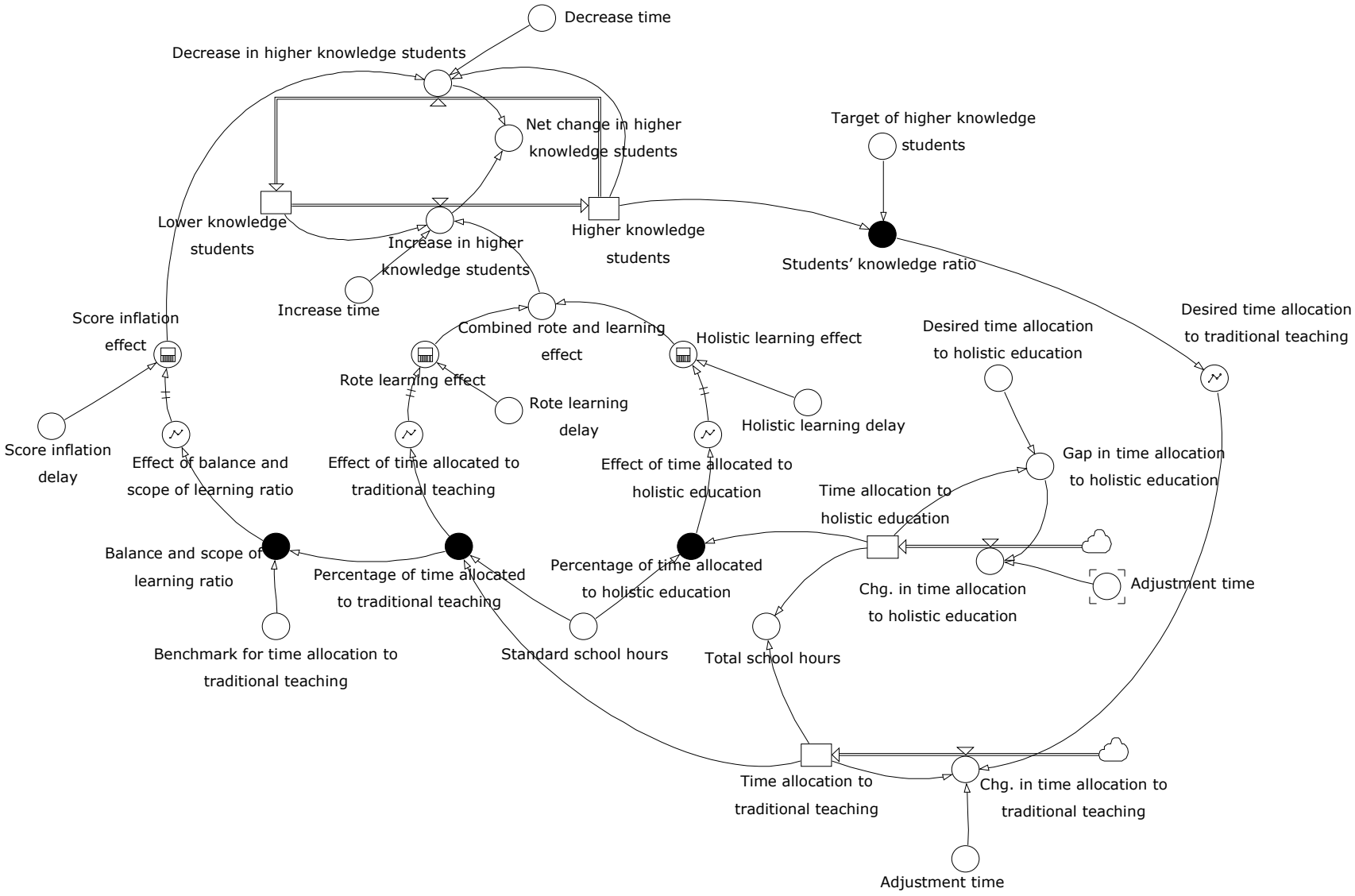


Figure 8: *A DPM model illustrating a policy based on the joint use of traditional and holistic education in Colombian public schools*

and outcomes from both traditional (i.e. notion-based) and holistic (i.e. project-based) education. While the former concerns the development of lower order thinking skills through rote learning and exam preparations, the latter refers to the development of higher order thinking skills (e.g. critical thinking, problem solving, teamwork aptitudes) that will allow students to solve real-life problems.

Therefore, the suggested policy will allow public schools to complement the teaching of tested subjects (e.g. math, languages, social and natural sciences) with the development of a deep learning in students through activities such as inter-disciplinary projects, group work, and student portfolios. Deep learning is the product from implementing a holistic education, which has been proved to be effective for enhancing performance in high-stake tests (Parker *et al.,* 2011; Scogin *et al.,* 2017; Ross *et al.,* 2001; Levine, 1994; Thomas, 2000).

In Figure 9 and Figure 10, the results from a simulation run of the extended model and an explanatory causal loop diagram are portrayed respectively. According to the simulation results, the public school can reduce its performance gap by increasing the time allocated to holistic education, which leads to raise the percentage of time allocated to holistic learning. As a consequence, the change in the number of higher knowledge students is affected positively. In the extended model, the "students' knowledge ratio" is significantly lower than the benchmark. Initially, this condition makes the school increase the time allocated to traditional teaching to enhance such a ratio (loop B-1 in Figure 10). As a result, a positive net change in the number of higher knowledge students is perceived during the first sixteen months of the simulation. To keep the results of the above policy in the long-term, the school must use overtime[1]. Therefore, the percentage of time allocated to holistic education (performance driver) must progressively be risen from 40% to 80% of standard school hours. As a result, a change in the loop dominance is perceived after sixteenth months because of an increase in the performance driver (loop B-2 in Figure 10).

This condition leads to reach a substantial stability in the net change of higher knowledge students and to sustain the gains from the short-term policy illustrated in loop-B1.

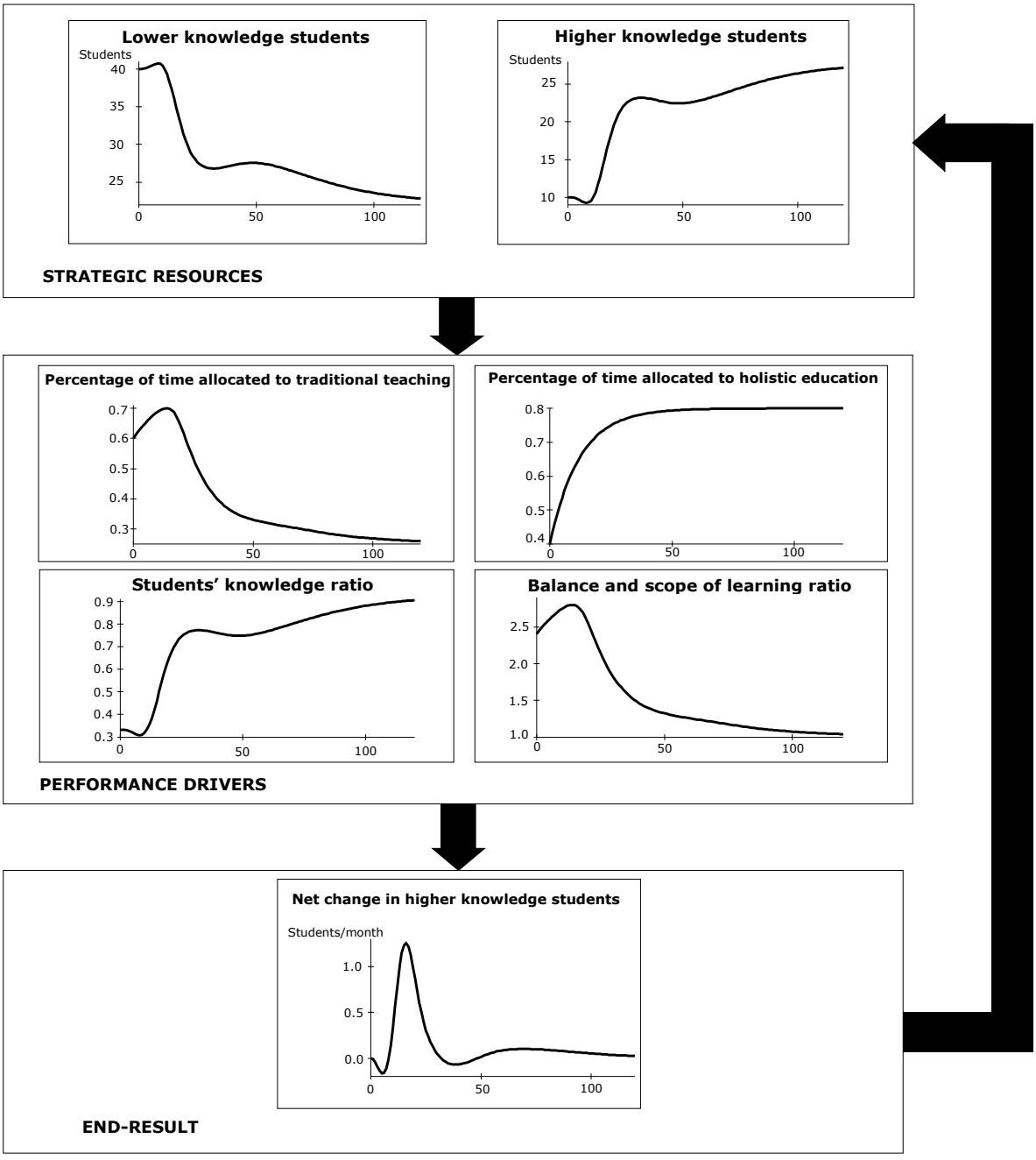


Figure 9: *Results from a simulation run illustrating the effects of combining traditional and holistic education in Colombian public schools*

---

[1] Colombian public schools often hire external contractors to train students for the standardized tests in out of school hours. In addition, regional governments and district education secretaries invest financial resources to develop extracurricular activities for improving students' performance in standardized tests. The educational policy, which is simulated by the extended model, implies the use of teacher overtime to foster a holistic education. This policy has been considered by most of the interviewees in this research as acceptable.

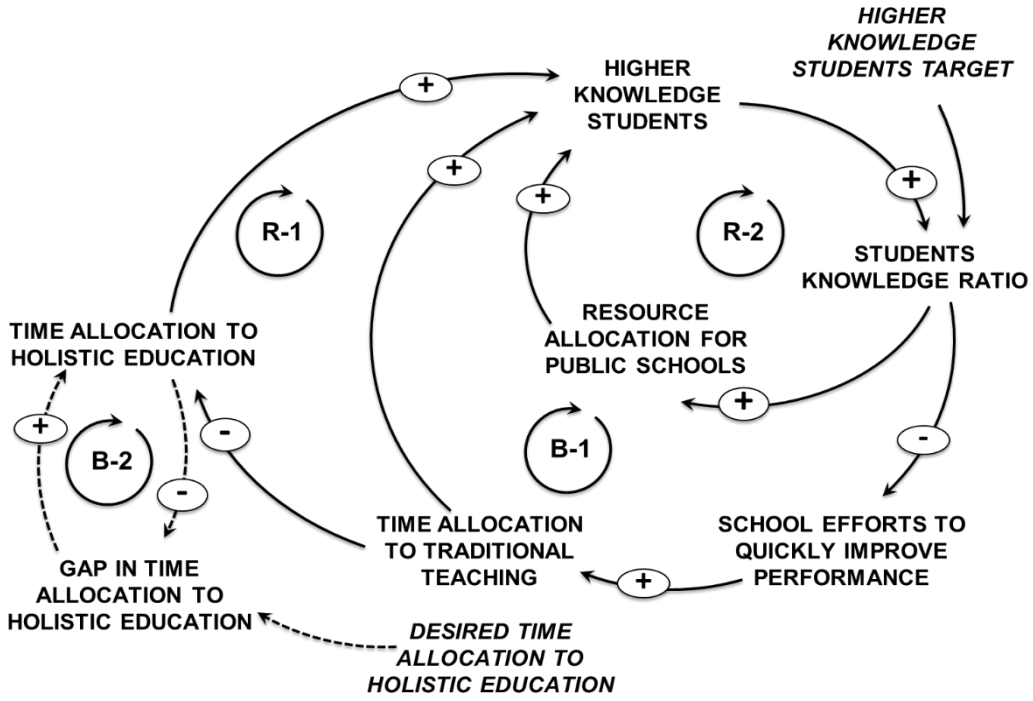


Figure 10: *Extending the scope of tests to holistic education in Colombian public schools*

Posteriorly, from the twenty-second month, the school reduces the time allocation to traditional teaching from 60% to 25% because of an enhancement in its performance. It makes the "Balance and scope of learning ratio" improve, which in turn slows down the loop "R1" in Figure 10 since the increase in lower knowledge students is reduced. As a result, the net change of higher knowledge students is positively affected in the long-term.

The above analysis shows how DPM can enhance the understanding of decision makers in complex and counterintuitive phenomena by identifying feedback-loops and dynamic performance measures (i.e. strategic resources, performance drivers, end-results). In particular, in the case-study of the Colombian public schools, an instrumental view of DPM has been used to help system designers and decision makers deal with behavioral distortions caused by the use of output-oriented performance measurement systems. Assessing school performance, through standardized tests, is a government mechanism to hold public schools accountable for student achievement in the short-term.

However, dysfunctional behaviors may emerge because of 1) an excessive pressure for increasing test scores, and 2) the adoption of inconsistent measurement systems. This condition may lead to a decline in school outcomes -such as

educational quality and social awareness- in the long-term. To avoid the previous side-effects, an extension in the boundaries of the adopted indicators has been proposed in this research by measuring the percentage of teaching time allocated to holistic education and its effect on school performance. As a result of this policy, an improvement in outputs and outcomes may be expected in the long-term.

In brief, the use of DPM to outline school assessment systems may foster a paradigm shift from an output to an outcome-based view in performance management. Such a view may support decision makers to counteract the effects of static and simplistic indicators intended to measure student learning in public schools.

5) CONCLUSIONS

5.1) Summary of the main discussions

Behavioral distortions from the use of school performance measurement systems have been discussed throughout this dissertation. An approach from general to specific has been used to develop the contents. The next lines summarize the main discussions of the previous chapters and sections.

In chapter 1, problem relevance, objective and research questions have been formulated. In particular, it has been discussed how the design of inconsistent performance measurement systems may trigger la emergence of behavioral distortions, which may produce a short-medium term increase in outputs and a long-term decrease in outcomes. Therefore, an approach aimed at fostering a shift from an output to an outcome-based view in performance management is required to prevent and counteract such dysfunctional behaviors. To this end, the instrumental view of Dynamic Performance Management (DPM) has been suggested in this research. Moreover, in order to set the conceptual roots on how to outline outcome-oriented performance measurement systems through DPM, the theory of street-level bureaucracies has been linked to the design of control systems through clan mechanisms.

In chapter 2, the first three research questions have been approached by an exhaustive literature review. This chapter has been divided into eight sections with several subsections in each of them. The contents of these sections are briefly described below.

In the first section, behavioral distortions from the use of performance measurement systems have been analyzed in a wide range of policy domains. In particular, the discussions of this section have been addressed by initially introducing broad concepts in performance measurement systems, intending to build basic notions for posterior analyses. Later, it has been discussed how performance indicators and several frameworks have been used in both public and private organizations to support decision-making. Posteriorly, criticisms and side-effects from the use of performance measurement systems in different policy settings have been debated. In addition, the relationship between the design and implementation

of output-oriented rewards, and the inclination of people to game the system has been explained. Finally, a broad spectrum of behavioral distortions in public schools have been discussed, and general recommendations to deal with them have been shared with the reader.

In the second section, this research has mainly been focused on the study of behavioral distortions in school contexts due to inconsistent assessment systems. To support this endeavor, at beginning of this section, basic notions in school accountability have been provided to the reader. Then, the history of test-based accountability systems, the use of PISA as an assessment tool to measure students' capacities for solving real-life problems, and the criticisms associated with the administration of standardized tests in public schools have been examined. Posteriorly, several recommendations to enhance school performance measurement systems and avoid common side-effects related to their adoption have been outlined by reviewing the preceding scientific literature.

In the third section, the concept of educational quality has been examined. Several opinions from experts in the field and organizations involved in the education sector have been portrayed. In addition, The Synthetic Education Quality Index, designed to measure the Colombian educational quality in public and private schools, has been explained. In particular, the four dimensions of performance (i.e. progress, performance, efficiency, school environment) -that are used to track student progress and address organizational efforts towards the attainment of school outcomes- have been discussed.

In the fourth section, the Colombian education system has been detailed. In particular, the organizational structure and the role of the most relevant internal units has been delineated. In addition, the academic levels and the legal framework, by which the education system is ruled in the country, have been described. Moreover, it has been explained why stakeholders in education (i.e. families, schools, government, and enterprises) should make coordinated efforts to achieve school outcomes in Colombia.

In the fifth section, relevant standardized tests, which are administered in Colombian public schools, have been examined. In particular, the history of the use of "SABER", "ICFES", "TIMSS" and "PISA" -designed to support the measurement of

student performance and the development of education policies in the Colombian education sector- has been described throughout this section. In addition, the use of summative and formative assessments to enhance the learning process of Colombian students has been discussed.

In the sixth section, opinions in favor and against the use of standardized tests to measure and track student achievement have been contrasted. Several scholars assert that standardized tests allow principals and schoolteachers to 1) know the competencies that students have gained over the years, and 2) develop school programs to reinforce strengths and mitigate weaknesses in the learning process. On the other hand, other scholars emphasize the detrimental effects that test-based accountability systems may trigger in the system, such as narrowing the school curricula to increase test scores. This section is concluded by depicting several recommendations on how to use standardized tests for reinforcing their advantages and minimizing their disadvantages.

In the seventh section, the design of control systems in organizations has been covered. Initially, the components of organizational control systems have been detailed. Posteriorly, an emphasis on the diagnostic control system and its relationship with the interactive control system has been denoted. In particular, it has been discussed the need of complementing the single-loop learning with the double-loop learning from the use of diagnostic and interactive control systems respectively. Then, the theory of street-level bureaucracy and the clan mechanism have been used to set the conceptual roots of an outcome-oriented view for outlining school performance measurement systems. At the end of this section, the need of implementing such view in the Colombian school context has been highlighted. To support this endeavor, Dynamic Performance Management (DPM) has been suggested.

In the eight section, the instrumental view of DPM has been explained by 1) outlining how performance drivers impact on outcome and output end-results, 2) determining how end-results affect strategic resources, and 3) understanding how strategic resources and benchmarks define the dynamics of performance drivers. This approach has also been used to foster a shift from an output- to an outcome-based view in performance management, intending to pursue sustainable outcomes

in the long-term. In the context of the Colombian public schools, DPM has been used to illustrate the weaknesses associated with the use of inconsistent performance measurement systems, and to support the design of a robust set of performance measures.

In chapters 3 and 4, the last two research questions have been approached. In particular, in chapter 3 the research methodology has been explained. First, the rationale for adopting DPM to support the design and implementation of consistent school performance measurement systems has been elucidated. Second, the strategy for collecting primary and secondary data has been detailed. Both sequential exploratory design and document review have been carried out in this investigation. Posteriorly, the approach to build and validate the DPM model -for the case-study of the Colombian public schools- has been illustrated. Finally, data outputs of this research have been specified.

In chapter 4, the modeling strategy and simulation results of the Colombian case-study have been discussed. Initially, a feedback view of behavioral distortions from the use of school assessment systems has been provided through a causal loop diagram (CLD). An emphasis on how dysfunctional behaviors may emerge as a result of perceived low performance ratios has been done. Posteriorly, problems with narrowing the curricula in Colombian public schools have been analyzed through a DPM chart and simulation runs from a DPM model. It implied that the adopted policies were examined by 2) distinguishing means from ends and identifying different “layers” of performance measures that captured contrasting time horizons, 2) framing and simulating feedback structures to understand complex and counterintuitive behaviors, and 3) extending the boundaries of action not only to single organizations but also to other relevant stakeholders. Finally, policy recommendations on how to outline a broader set of performance measures and benchmarks to prevent school behavioral distortions have been delineated. To this end, an outcome-oriented view in performance management has been used. In particular, an educational policy -that complements the teaching of disciplines assessed through standardized tests (e.g. math, languages, natural and social sciences) with the development of student skills and attitudes through holistic activities (e.g. inter-disciplinary projects, group work, and student portfolios)- has been proposed at the end of this chapter.

5.2) Contribution to the existing knowledge

This thesis has been aimed at contributing to the domain of the behavioral distortions associated with the use of school performance measurement systems. In particular, it has been shown how an inconsistent design of test-based accountability systems may jeopardize the attainment of school outcomes such as educational quality. The case-study of the Colombian public schools has been used to show how an outcome-based view in performance management can be useful to 1) challenge the consistency of the adopted indicators, and 2) deal with behavioral distortions, such as narrowing the school curricula to improve performance in standardized tests. Dynamic Performance Management (DPM) has been suggested to prevent and counteract such dysfunctional behaviors by understanding how policy levers and performance drivers impact on outputs and outcomes, and how such end-results feedback into strategic resources.

The case-study has shown how both an excessive government pressure for improving performance in high-stake tests and the adoption of output-oriented reward systems have provoked the emergence of gaming behaviors in Colombian public schools (e.g. narrowing the curricula). In particular, an increase in the time allocated to traditional teaching to the detriment of a holistic education has been observed. Therefore, an extension of the system boundaries, from an output to an outcome-based view, is crucial to mitigate negative behavioral effects of using standardized tests to hold public schools accountable for student performance. In addition, the case-study has been used to highlight the need of 1) taking into account the socio-economic background where public schools operate, and 2) outlining a common shared view among distinct stakeholders in education, intending to pursue sustainable outcomes in the long-term. To this end, DPM has also been used to set the basis for policy discussion by filtering the insights from modeling and simulation through the lenses of decision makers in public schools.

Finally, this research has suggested combining both traditional and holistic education to enhance the achievement of school outcomes. This approach implies that 1) a proper development of basic notions in different knowledge areas is a prerequisite for building higher order thinking skills in school students, and 2) the use of standardized tests to assess student performance is not bad per se because such

tests can support pedagogical processes. However, decision makers should know their limitations to maximize the benefits of using them. Therefore, test results should not be conceived as an absolute truth for policy making, but as a tool to enhance teaching strategies. The results from this research also suggest that an increase in test scores does not imply an increase in educational quality. However, it would be expected that an increase in educational quality would lead to an increase in test scores.

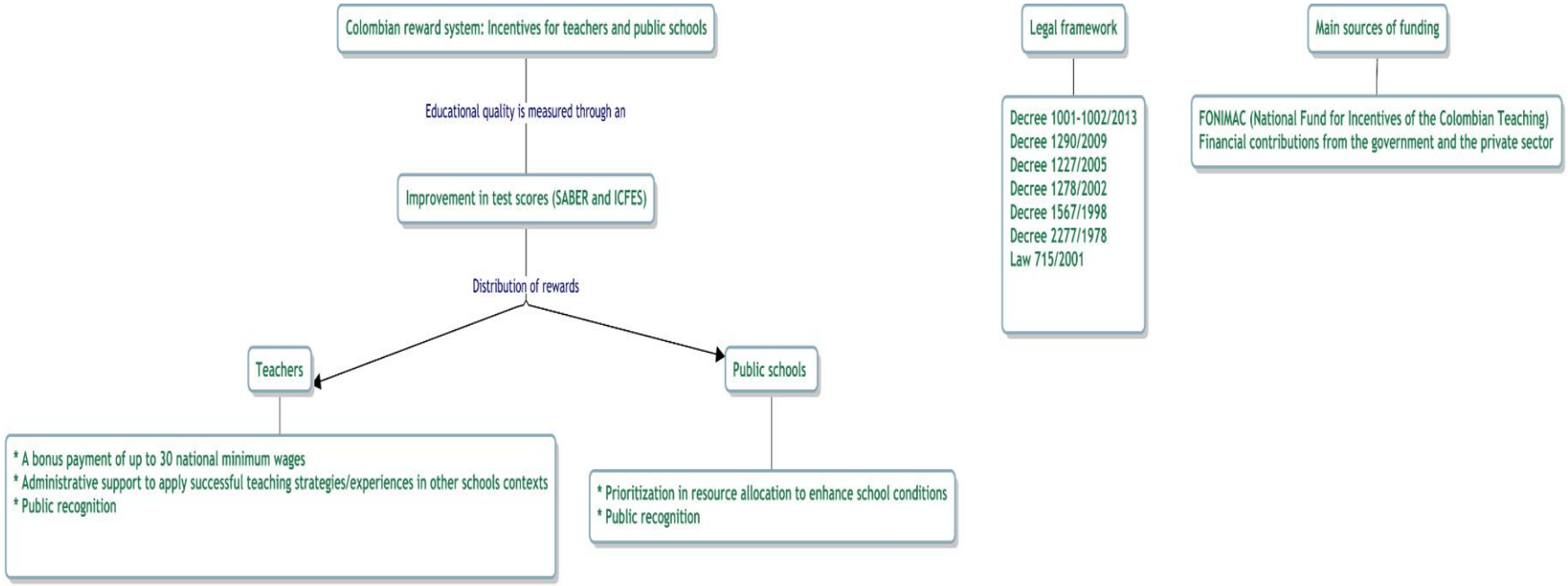
5.3) Limitations and indications for further research

The results should not be made generalizable because it was used a small sample size of schoolteachers to collect the primary data. Instead, this thesis should be seen as an initial step to analyze behavioral distortions in public schools through an outcome-oriented perspective in performance management. To this end, Dynamic Performance Management (DPM) has been implemented in the case-study of the Colombian public schools. As a result, a DPM simulation model has been built to prevent and counteract dysfunctional behaviors caused by the adoption of inconsistent performance measures. Such DPM model could be enhanced by 1) collecting additional data from the results of implementing the current standardized test system in Colombia, and 2) working in a multi-stakeholder setting to identify shared strategic resources and outline sustainable school policies through a collaborative perspective. This implies the need of broadening the focus of performance evaluation from test scores to the assessment of school and community outcomes (e.g. educational quality, social capital, civic-mindedness, social awareness, attractiveness of the local area, trust in government).

Therefore, the development of further DPM models and case-studies -from different social contexts- will be needed to extend the boundaries of the research. This may lead to incorporate other important variables that may affect school performance (e.g. level of holistic skills).

APPENDICES

A) Colombian reward system: Incentives for teachers and public schools



B) First survey

UNIVERSITÀ DEGLI STUDI DI PALERMO

This survey is part of the doctoral thesis conducted by the student Robinson Stevens Salazar Rua, who is enrolled in the program "Model Based Public Planning, Policy Design and Management" at the University of Palermo in Italy. The purpose of this survey is to determine how stakeholders in education impact on educational quality. To this end, the case of the Colombian public schools is analyzed through the lenses of a Dynamic Performance Management (DPM) approach.

In order to collect the primary data of this research, this survey, based on open-ended questions, was designed. Participants are requested to analyze each of the questions in this document carefully.

The information provided by the participants will be used for research purposes in this doctoral thesis and in academic publications. Participants are free to write their names, phone number and e-mail to be interviewed. The interviews will be recorded if the participants agree with it.

Should you have questions, please contact me by the following e-mail address: rssalazarr@gmail.com

Thanks for your participation.

Participant details

Name:

Current job position:

Years of work experience in the education sector:

Phone number/e-mail for the interview:

Date:

Note: Answers to the questions must be provided in order. **This survey must be completed before December 15, 2016.**

Questions

1) How do you define the concept *"Educational quality"?*

2) What are your duties and responsibilities to improve student achievement in public schools?

3) What should the role of stakeholders (i.e. schools, government, families, enterprises) be to improve educational quality? Do you think a gap between role expectations and role performance exists in the Colombian education system? Why?

4) Do you think schoolteachers' role is properly valued in Colombian society? If the answer is negative, do you think an undervalued perception of teaching activities may impact on educational quality? Why?

5) What is your opinion about using standardized tests to measure educational quality? Do you think it is an appropriate approach? If the answer is negative, what do you suggest for tracking student progress and school performance?

6) Which strategies have you ever used to improve student performance in standardized tests? Do you think such strategies also contribute enhancing the achievement of school outcomes? Have you ever observed unintended results of implementing such strategies in the medium and long-term?

7) In the Colombian context, what are the main outcomes that might be attained if school performance is enhanced through a consistent design of education policies?

C) Second survey

UNIVERSITÀ DEGLI STUDI DI PALERMO

This survey is part of the doctoral thesis conducted by the student Robinson Stevens Salazar Rua, who is enrolled in the program "Model Based Public Planning, Policy Design and Management" at the University of Palermo in Italy. This survey has a dual purpose. In the first instance, it is aimed at understanding how standardized tests impact on the behavior of Colombian schoolteachers. In the second instance, it is aimed at validating the structure of a Dynamic Performance Management (DPM) model, which has been built to explain and counteract the negative effects of school practices aimed at increasing test scores.

In order to collect the primary data of this research, this survey, based on open-ended questions and numerical reasoning, was designed. Participants are requested to analyze each of the questions in this document carefully.

The information provided by the participants will be used for research purposes in this doctoral thesis and in academic publications. Participants are free to write their names, phone number and e-mail to be interviewed. The interviews will be recorded if the participants agree with it.

Should you have questions, please contact me by the following e-mail address: rssalazarr@gmail.com

Thanks for your participation.

Participant details

Name:

Current job position:

Years of work experience in the education sector:

Phone number/e-mail for the interview:

Date:

Note: Answers to the questions must be provided in order. **This survey must be completed before January 10, 2019.**

Questions

1) What kind of education do you consider the most appropriate to build higher order thinking skills (e.g. critical thinking, problem-solving, and teamwork)? Why?

2) Based on your experience, how long does it take a student to gain an adequate level of higher order thinking skills?

3) In an "ideal" school, how much time of the school year should be devoted to activities that look for developing higher order thinking skills? Why?

4) Do you think the current standardized tests are aligned with an education that fosters the development of higher order thinking skills? Why?

5) How much time of the school year does your school allocate to exam preparations? Is it possible to increase/decrease such time allocation under certain circumstances? If the answer is positive, could you quantify it?

6) Do you think school strategies intended to develop higher order thinking skills may impact on test scores? How?

7) Do you think resource investments to improve performance in standardized tests can also enhance the development of higher order thinking skills? How?

8) Based on your experience, how much time does a public school require improving its performance in standardized tests significantly?

9) How often does your school adjust its policies and strategies?

10) What are the effects of prioritizing traditional education (i.e. teaching aimed at building basic notions in students through rote learning activities and exam preparations) over holistic education (i.e. teaching aimed at building higher order thinking skills)?

11) Which of the following options do you consider the closest one to reality? Why? (Note: Read the statement carefully and underline the option in italics that is most appropriate)

- The time to increase the number of lower knowledge students to higher knowledge students is *greater than / equal to / less than* the time to decrease the number of higher knowledge students to lower knowledge students.

12) Could you quantify the previous statement?

- Example: I think the time to increase the number of lower knowledge students to higher knowledge students is 5 times *greater than / less than* the time to decrease the number of higher knowledge students to lower knowledge students.

Graphs

You are requested to plot X vs. Y graphs in this section, intending to determine the relationship between the variables (Note: In all the statements, the first variable corresponds to the X-axis and the second variable corresponds to the Y-axis).

a) Students knowledge ratio (Higher knowledge students over Target of higher knowledge students) and Desired time allocation to traditional teaching

b) Percentage of time allocated to traditional teaching and its effect on the increase in higher knowledge students

c) Percentage of time allocated to traditional teaching and its effect on the increase in lower knowledge students

d) Percentage of time allocated to holistic education and its effect on the increase in higher knowledge students

D) Model documentation

1) DPM model - Initial version

Stocks and flows

Higher knowledge students (t) = Higher knowledge students (t - dt) + (Increase in higher knowledge students - Decrease in higher knowledge students) * dt INIT = 10 (students)

Inflow:

Increase in higher knowledge students = ('Lower knowledge students' / 'Increase time') * 'Rote effect'

Outflow:

Decrease in higher knowledge students= ('Higher knowledge students' / 'Decrease time') * 'Score inflation effect'

Lower knowledge students (t) = Lower knowledge students (t - dt) + (Decrease in higher knowledge students - Increase in higher knowledge students) * dt INIT = 40 (students)

Inflow:

Decrease in higher knowledge students = ('Higher knowledge students' / 'Decrease time') * 'Score inflation effect'

Outflow:

Increase in higher knowledge students = ('Lower knowledge students' / 'Increase time') * 'Rote effect'

Teaching time allocation to tested subjects (t) = Teaching time allocation to tested subjects (t - dt) + (Chg in teaching time allocation to tested subjects) * dt INIT = 0.25 (dimensionless)

Inflow and Outflow:

Chg in teaching time allocation to tested subjects = ('Desired teaching time allocation to tested subjects' - 'Teaching time allocation to tested subjects') / 'Adjustment time'

Auxiliaries and parameters

Net change in higher knowledge students= 'Increase in higher knowledge students' - 'Decrease in higher knowledge students'

Students knowledge ratio= 'Higher knowledge students' / 'Target of higher knowledge students'

Percentage of time allocated to tested subjects= 'Teaching time allocation to tested subjects' / 'Total school hours'

Balance and scope of learning ratio= 'Percentage of time allocated to tested subjects' / 'Benchmark for time allocation to tested subjects'

Score inflation effect= DELAYINF('Effect of balance and scope of learning ratio','Score inflation delay',2,0)

Rote effect= DELAYINF('Effect of time allocated to tested subjects','Rote delay',2,0)

Effect of balance and scope of learning ratio= GRAPH('Balance and scope of learning ratio',1,0.2,{1,1.007,1.052,1.225,1.509,1.861,2.213,2.610,2.865,2.985,3//Min:1;Max:3//}

Effect of time allocated to tested subjects= GRAPH('Percentage of time allocated to tested subjects',0.25,0.055,{1,1.022,1.086,1.206,1.367,1.558,1.708,1.839,1.929,1.978,1.993//Min:1;Max:2//})

Desired teaching time allocation to tested subjects= GRAPH('Students knowledge ratio',0,0.0833,{0.8,0.798,0.792,0.784,0.749,0.703,0.633,0.499,0.367,0.304,0.271,0.254,0.250//Min:0.25;Max:0.8//})

Adjustment time = 12 (months)

Score inflation delay= 36 (months)

Rote delay=12 (months)

Increase time = 24 (months)

Decrease time=3 (months)

Target of higher knowledge students = 45 (students)

Benchmark for time allocation to tested subjects = 0.25 (dimensionless)

Total school hours = 1 (dimensionless)

2) DPM model - Extended version

Stocks and flows

Higher knowledge students (t) = Higher knowledge students (t - dt) + (Increase in higher knowledge students - Decrease in higher knowledge students) * dt INIT = 10 (students)

Inflow:

Increase in higher knowledge students= ('Lower knowledge students' / 'Increase time') * 'Combined rote and learning effect'

Outflow:

Decrease in higher knowledge students= ('Higher knowledge students' / 'Decrease time') * 'Score inflation effect'

Lower knowledge students (t) = Lower knowledge students (t - dt) + (Decrease in higher knowledge students - Increase in higher knowledge students) * dt INIT = 40 (students)

Inflow:

Decrease in higher knowledge students= ('Higher knowledge students' / 'Decrease time') * 'Score inflation effect'

Outflow:

Increase in higher knowledge students= ('Lower knowledge students' / 'Increase time') * 'Combined rote and learning effect'

Time allocation to holistic education (t) = Time allocation to holistic education (t - dt) + (Chg in time allocation to holistic education) * dt INIT = 0.4 (dimensionless)

Inflow and Outflow:

Chg in time allocation to holistic education= 'Gap in time allocation to holistic education' / 'Adjustment time'

Time allocation to traditional teaching(t) = Time allocation to traditional teaching (t - dt) + (Chg in time allocation to traditional teaching) * dt INIT = 0.6 (dimensionless)

Inflow and outflow:

Chg in time allocation to traditional teaching= ('Desired time allocation to traditional teaching' - 'Time allocation to traditional teaching') / 'Adjustment time'

Auxiliaries and parameters

Total school hours= 'Time allocation to holistic education' + 'Time allocation to traditional teaching'

Percentage of time allocated to holistic education= 'Time allocation to holistic education' / 'Standard school hours'

Percentage of time allocated to traditional teaching= 'Time allocation to traditional teaching' / 'Standard school hours'

Balance and scope of learning ratio= 'Percentage of time allocated to traditional teaching' / 'Benchmark for time allocation to traditional teaching'

Combined rote and learning effect= 'Rote learning effect' * 'Holistic learning effect'

Net change in higher knowledge students= 'Increase in higher knowledge students' - 'Decrease in higher knowledge students'

Gap in time allocation to holistic education= 'Desired time allocation to holistic education' - 'Time allocation to holistic education'

Students knowledge ratio= 'Higher knowledge students' / 'Target of higher knowledge students'

Score inflation effect= DELAYINF('Effect of balance and scope of learning ratio','Score inflation delay',2,0)

Rote learning effect= DELAYINF('Effect of time allocated to traditional teaching','Rote learning delay',2,0)

Holistic learning effect= DELAYINF('Effect of time allocated to holistic education', 'Holistic learning delay',2,0)

Effect of balance and scope of learning ratio= GRAPH('Balance and scope of learning ratio',1,0.2,{1,1.007,1.052,1.225,1.509,1.861,2.213,2.610,2.865,2.985,3//Min:1;Max:3//})

Effect of time allocated to traditional teaching= GRAPH('Percentage of time allocated to traditionalteaching',0.25,0.055,{1,1.022,1.086,1.206,1.367,1.558,1.708,1.839,1.929,1.978, 1.993//Min:1;Max:2//})

Effect of time allocated to holistic education= GRAPH('Percentage of time allocated to holisticeducation',0.25,0.055,{0.2,0.6,1.2,1.9,3.13,4.54,6.34,8.1,9.12,9.58,9.96//Min:0;Max: 10//})

Desired time allocation to traditional teaching= GRAPH('Students knowledge ratio',0,0.0833,{0.8,0.798,0.792,0.784,0.749,0.703,0.633,0.499,0.367,0.304,0.271,0.254,0.25 0//Min:0.25;Max:0.8//})

Adjustment time=12 (months)

Score inflation delay= 36 (months)

Rote learning delay= 12 (months)

Holistic learning delay= 24 (months)

Increase time= 24 (months)

Decrease time= 3 (months)

Target of higher knowledge students= 30 (students)

Desired time allocation to holistic education= 0.8 (dimensionless)

Standard school hours= 1 (dimensionless)

Benchmark for time allocation to traditional teaching= 0.25 (dimensionless)

REFERENCES

Adab P, Rouse AM, Mohammed MA, Marshall T. 2002. *Performance league tables: the NHS deserves better*. BMJ. 324(7329): 95-98.

Adams, JE, Kirst M. 1999. *New Demands for Educational Accountability: Striving for Results in an Era of Excellence.* In Murphy J and Louis KS (eds.), Handbook of Research in Educational Administration, San Francisco, CA: Jossey-Bass.

Adler PS, Borys B. 1996. *Two types of bureaucracy: Enabling and coercive*. Administrative Science Quarterly. 41(1): 61–89.

Ahrens T, Chapman CS. 2004. *Accounting for flexibility and efficiency: A field study of management control systems in a restaurant chain.* Contemporary Accounting Research. 21(2): 271–301.

Airasian PW. 1987. *State mandated testing and educational re-form: Context and consequences*. American Journal of Education. 95: 393–412.

Allington RL, McGill-Franzen A. 1992. *Unintended effects of educational reform in NewYork*. Educational Policy. 6(4): 397–414.

Altrichter H, Kemethofer D. 2015. *Does accountability pressure through school inspections promote school improvement?* School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice. 26(1): 32-56. DOI: 10.1080/09243453.2014.927369

Andersen DL, Luna-Reyes LF, Diker VG, Black L, Rich E, Andersen DF. 2012. *The disconfirmatory interview as a strategy for the assessment of system dynamics models*. System Dynamics Review. 28: 255-275.

Argyris C. 1952. *The Impact of Budgets on People*. School of Business and Public Administration, Cornell University, Ithaca.

Argyris C. 1977. *Double Loop Learning in Organizations*. Harvard Business Review, September-October 1977. 115-124.

Austin R, Gittel JH. 2002. *Anomalies of measurement: when it works but should not. Business performance measurement: theory and practice.* A. Neely. Cambridge, Cambridge University Press: xiii, 366.

Ballou D. 2001. *Pay for performance in public and private schools*. Economics of Education Review. 20: 51-61. Retrieved from http://dx.doi.org/10.1016/S0272-7757(99)00060-6

Barlas Y. 1996. *Formal aspects of model validity and validation in system dynamics*. System Dynamics Review. 12: 183-210.

Berliner JS, 1956. *A problem in Soviet business management*. Administrative Science Quarterly. 1: 86-101.

Berman E. 2002. *How useful is performance measurement*? Public Performance & Management Review. 25(4): 348-351.

Bianchi C. 2010. *Improving performance and fostering accountability in the public sector through system dynamics modeling: From an "external" to an "internal" perspective*. Systems Research and Behavioral Science. 27: 361–384.

Bianchi C. 2012. *Enhancing performance management and sustainable organizational growth through system-dynamics modelling. Systemic management for intelligent organizations.* Berlin, Germany: Springer. 143-161.

Bianchi C. 2016. *Dynamic Performance Management*. Springer International Publishing, Zurich, Switzerland.

Bianchi C, Bovaird T, Loeffler E. 2017. *Applying a Dynamic Performance Management Framework to Wicked Issues: How Coproduction Helps to Transform Young People's Services in Surrey County Council, UK. International Journal of Public Administration* 40, 10, 833-846. doi:10.1080/01900692.2017.1280822

Bianchi C, Rivenbark W. 2014. *Performance Management in Local Government: The Application of System Dynamics to Promote Data Use* (with W. Rivenbark), International Journal of Public Administration, 37.

Bianchi C, Salazar Rua RS. 2017. *Applying Dynamic Performance Management to detect behavioral distortions associated with the use of formal performance measurement systems in public schools: the case of Colombia*. Paper presented at the APPAM Conference, Chicago.

Bianchi C, Williams DW. 2015. *Applying System Dynamics Modeling to foster a cause and effect perspective in dealing with behavioral distortions associated with a city's performance measurement programs*. Public Performance & Management. 395-425.

Bianchi C, Winch GW. 2006. *Unleashing growth potential in 'stunted' SMEs: Insights from simulation experiments.* International Journal of Entrepreneurship and Small Business. 3(1): 92–105.

Bird SM, Cox D, Farewell VT, Goldstein H, Holt T, Smith PC. 2005. *Performance indicators: Good, bad, and ugly.* Journal of the Royal Statistical Society: Series A (Statistics in Society). 168(1): 1–27.

Birnberg JG, Turopolec L, Young SM. 1983. *The Organizational Context of Accounting*. Accounting, Organizations and Society. 8(2): 111-129.

Blau PM, Meyer MW. 1971. *Bureaucracy in Modern Society*. 2nd Edition. Random House, New York.

Bohte J, Meier K. 2000. *Goal Displacement: Assessing the Motivation for Organizational Cheating.* Public Administration Review. 60(2): 173 – 182.

Booher-Jennings J. 2007. *Michael Fullan, Turnaround Leadership*. Book Review, Journal of Educational Change. 8(3): 291-294.

Borgonovi E. 1996. *Principi e sistemi aziendali per le Amministrazioni Pubbliche (Management Principles and Systems for Public Administrations)*. Egea, Milano.

Bosker RJ, Scheerens J. 2000. *Publishing school performance data*. European Education. 32(3): 12-30.

Bourne M, Franco M. 2003. *Corporate Performance Management.* SAS Institute Inc. Marlow, Centre for Business Performance, Cranfield University School of Management.

Bourne M, Neely A, Platts K, Mills J. 2002. *The success and failure of performance measurement initiatives. Perceptions of participating managers.* International Journal of Operations and Production Management. 22(11).

Bracci E. 2009. *Autonomy, Responsibility and Accountability in the Italian School System.* Critical Perspectives on Accounting. 20(3): 293-312.

Breakspear S. 2014. *How does pisa shape education policy making?How does PISA shape education policy making? Why how we measure learning determines what counts in education.* Centre for Strategic Education.

Brenner B. 2009. *Management Control in Central and Eastern European Subsidiaries*.

Brigham BH, Fitzgerald L. 2001. *Controlling Managers and Organisations: The Case of Performance Measurement in a Regulated Water Company.* Centre for Management under Regulation, Warwick Business School, Warwick.

Brimblecombe N, Ormston M, Shaw M. 1996. *Teachers' perceptions of inspections*. In J. Ouston, P. Earley, & B. Fidler (eds.), Ofsted inspections: The early experience (pp. 126–134). London: David Fulton Publishers.

Brodkin EZ. 2008. *Accountability in Street-Level Organizations.* Journal of Public Administration. 31(3): 317-336.

Brodkin EZ. 2011. *Policy Work: Street-Level Organizations Under New Managerialism.* Journal of Public Administration Research and Theory. 21(2): i253-i277.

Brock D, Powell M, Hinings CR (eds). 1999. *Restructuring the Professional Organisation: Accounting, Healthcare, and Law*. London: Routledge.

Brown A. 2005. *Implementing performance management in England's primary schools*. International Journal of Productivity and Performance Management. 54(5/6): 468-481.

Buchanan B. 1974. *Government Managers, Business Executives, and Organizational Commitment.* Public Administration Review. 35: 339-347.

Cajiao F. 2008. *Diálogo nacional sobre la evaluación del aprendizaje en el aula [National dialogue about the learning assessment in the classroom]*. Bogota, Colombia: Ministry of education.

Campbell DT. 1969. *Reforms as experiments*. American Psychologist. 24(4): 35.

Camperos M. 1984. *La evaluación formativa del aprendizaje [The formative evaluation of the learning]*. Faculty of humanities. Caracas, Venezuela: Mimeo.

Carnoy M. 2001. *El impacto de la mundialización en las estrategias de reforma educativa. [The impact of globalization on education reform strategies]* en Revista de Educación, n.o extra, págs. 101-110.

Carnoy M, Elmore R, Siskin L (eds). 2003. *The New Accountability: High Schools and High Stakes Testing*. New York, NY: RoutledgeFalmer.

Case, P, Case, S, Catling, S. 2000. *Please show you're working; A critical assessment of the impact of Ofsted inspection on primary teachers.* British Journal of Sociology of Education. 21: 605–621.

Castellano JF, Young S, Roehm HA. 2004. *The seven fatal flaws of Performance Measurement*. The CPA Journal. 74(6): 32.

Center on Education Policy. 2009. *State test score trends through 2007-08, part 2: Is there a plateau effect in test scores?* Washington, DC: Author. Retrieved from https://files.eric.ed.gov/fulltext/ED506122.pdf

Chapman C. 2001. *Changing classrooms through inspections*. School Leadership and Management. 21: 59–73.

Clarke M, Haney W, Madaus G. 2000. *High stakes testing and high school completion*. National Board on Educational Testing and Public Policy Statements.

Clotfelter CT, Ladd H. 1996. *Recognising and rewarding success in public schools*. In Ladd HF ed. Holding Schools Accountable: Performance Based Reform in Education. Brookings, Washington DC. 23-64.

Codesocial. 2009. *Organización del sistema educativo: Conceptos generales de la educación preescolar, básica y media [Organization of the Education System: General concepts of the pre-school, basic and vocational education]*. Bogota, Colombia: Ministry of education.

Collins S, Davis-Molin W, Conley D. 2013. *Journey toward deeper learning: An evaluation of the Roadtrip Nation Experience in the San Jose PLUS academies*. Eugene, OR: Educational Policy Improvement Center.

Colombian ministry of education. 2013. *Colombia en PISA 2012: Principales resultados [Colombia in PISA 2012: Main results]*. Bogota, Colombia: Ministry of education.

Courty P, Marschke G. 2003. *Dynamics of performance-measurement systems*. Oxford Review of Economic Policy. 19(2): 268–284.

Courty P, Marschke G. 2004. *An empirical investigation of gaming responses to explicit performance incentives*. Journal of Labor Economics. 22(1): 23–56.

Courty P, Marschke G. 2007. *Making government accountable: Lessons from a federal job training program*. Public Administration Review. 67(5): 904–916.

Cuganesan S, Guthrie J, Vranic V. 2014. *The riskiness of public sector performance measurement: A review and research agenda*. Financial Accountability & Management. 30(3): 279–302.

Cullen JB, Reback R. 2006. *Tinkering toward accolades: School gaming under a performance accountability system*. In T. Gronberg & D. Jansen (eds.). Advances in applied microeconomics (14). Improving school accountability: Check-ups or choice (pp. 1–34). Amsterdam: Elsevier Science.

Cullingford C, Daniels S. 1999. *Effects of OFSTED inspections on school performance*. In Cullingford C (Ed.), An inspector calls (pp. 59–69). London: Kogan Page.

Darling-Hammond L. 1991. *The implications of testing policy for quality and equality*. Phi Delta Kappan. 73(3): 220–225.

Darling-Hammond L. 2004. *Standards, Accountability, and School Reform.* Teachers College Record. 106(6): 1047-1085.

Darling-Hammond L, McCloskey L. 2008. *Assessment for Learning around the World, What Would it Mean to Be Internationally Competitive?* Phi Delta Kappan. 90(4): 263–272.

Darrow AA. 2016. *The Every Student Succeeds Act (ESSA): What It Means for Students With Disabilities and Music Educators*. General Music Today.

Davies H, Mannion R, Goddard M, Smith PC. 2000. *How Health Care Providers Use Comparative Outcomes Data to Monitor and Improve Care: A Cross-National Study*. Paper presented at the Academy for Health Services Research and Health Policy Meeting.

Deci EL. 1971. *Effects of externally mediated rewards on intrinsic motivation*. Journal of Personality and Social Psychology. 18: 105-115.

Deci EL. 1972. *The effects of contingent and noncontingent rewards and controls on intrinsic motivation*. Organizational Behavior and Human Performance. 18: 217-229.

Deci EL. 1976. *The Hidden Costs of Rewards*. Organizational Dynamics. 4(3): 61-72.

Deci EL, Betlly G, Kahle J, Abrams L, Porac J. 1981. *When trying to win: Competition and intrinsic motivation.* Personality and Social Psychology Bulletin. 7: 79-83.

Deci EL, Koestner R, Ryan RM. 2001. *Extrinsic Rewards and Intrinsic Motivation in Education: Reconsidered Once Again*. Review of Educational Research. 71(1): 1-27.

Deci EL, Koestner R, Ryan RM. 1999. *A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation*. Psychological Bulletin. 125: 627–668.

Deere D, Strayer W. 2001. *Putting schools to the test: school accountability, incentives and behavior.* Working paper, Department of Economics, Texas A&M University.

De Lancer Julnes P. 2006. *Performance Measurement: An Effective Tool for Government Accountability? The Debate Goes On*. Evaluation. 12(2): 219–235.

Deming WE. 1986. *Out of the Crisis: The MIT Press*.

De Waal AA. 2004. *Stimulating performance-driven behaviour to obtain better results.* International Journal of Productivity and Performance Management. 53(4): 301 - 316.

De Wolf IF, Janssens FJG. 2007. *Effects and side effects of inspections and accountability in education: An overview of empirical studies*. Oxford Review of Education. 33: 379–396.

Downs A. 1967. *Inside Bureaucracy*. Little, Brown, Boston.

Duncombe W, Yinger J. 1998. *School finance reform: Aid formulas and equity objectives.* The National Tax Journal. 51(2): 239-62.

Ehren MCM. 2006. *Toezicht en schoolverbetering* [School inspection and school improvement] (PhD dissertation).

Elmore RF. 1979. *Backward Mapping: Implementation Research and Policy Decisions.* Political science quarterly. 94(4): 601–616.

Elmore RF, Abelman CH, Fuhrman SH. 1996. *The new accountability in state education reform: from process to performance.* In Ladd, H.F. ed Holding Schools Accountable: Performance Based Reform in Education. Brookings, Washington DC. 23-64.

EPPI. 2002. *A Systematic Review of the Impact of Summative Assessment and Tests on Students' Motivation for Learning*. London: EPPI-Centre, Institute of Education, University of London.

Escudero Muñoz JM. 2003. *La calidad de la educación: Controversias y retos para la Educación Pública [Educational quality: Controversy and challenges for the public education]*. Murcia, Spain: University of Murcia.

Evans JR. 2004. *An exploratory study of performance measurement systems and relationships with performance results*. Journal of Operations Management. 22(3): 219-232.

Fernández Gómez HG. 2005. *Cómo interpretar la evaluación de las pruebas SABER. [How to interpret the SABER standardized exam]*. Bogota, Colombia: Ministry of education.

Figlio D. 2002a. *Aggregation and accountability*. In C. Finn, (Ed.) No child left behind: What will it take? Washington, D.C.: Thomas B. Fordham Foundation.

Figlio D. 2005. *Measuring school performance: Promise and pitfalls.* In L. Stiefel, AM. Schwartz, R. Rubenstein, J. Zabel (eds.). Measuring School Performance and Efficiency: Implications for practice and research, 2005. Yearbook of the American Education Finance Association. Larchmont: NY, Eye on Education.

Figlio D, Getzer LS. 2002, April. *Accountability, ability, and disability: Gaming the system?* Cambridge, MA: National Bureau of Economic Research.

Figlio D, Getzler LS. 2006. *Accountability, ability and disability: Gaming the system*. In T. Gronberg & D. Jansen (eds.). Advances in applied microeconomics (14). Improving school accountability: Check- ups or choice (pp. 35–49). Amsterdam: Elsevier Science.

Figlio D, Rouse C. 2004. *Do accountability and voucher threats improve low-performing schools?* (Working paper). Cambridge, MA: National Bureau of Economic Research.

Figlio D, Rouse C. 2006. *Do accountability and voucher threats improve low-performing schools?* Journal of Public Economics. 90(1): 239–255.

Finnish National Board of Education. 2004. *National core curriculum for basic education 2004.* Retrieved from https://www.oph.fi/sites/default/files/documents/perusopetuksen-opetussuunnitelman-perusteet_2004.pdf

Firestone W. 2002. *Educational Accountability.* Encyclopedia of Education.

Firestone WA, Pennell JR. 1993. *Teacher commitment, working conditions, and differential incentive policies*. Review of Educational Research. 63: 489-525. doi:10.3102/00346543063004489

Fitz-Gibbon C, Stephenson-Forster N. 1999. *Is Ofsted helpful? An evaluation using social science criteria*. In C. Cullingford (Ed.), An inspector calls: Ofsted and its effect on school standards (pp. 97–118). London: Kogan Page.

Flamholtz EG. 1983. *Accounting, budgeting and control systems in their organizational context: Theoretical and empirical perspectives.* Accounting, Organizations and Society. 8(2/3): 153-169.

Forester T. 1993. *Silicon Samurai: How Japan conquered the world's IT industry*. Blackwell Business.

Frey BS, Benz M. 2005. *Can private learn from public governance?* The Economic Journal. 115(507): F377–F396.

Friedlaender D, Burns D, Lewis-Charp H, Cook-Harvey CM, Darling-Hammond L. 2014. *Student-centered schools: Closing the opportunity gap*. Palo Alto, CA: Stanford Center for Opportunity Policy in Education.

Fuller B. 2004, August 22. *Accountability rises, scores fall.* Los Angeles Times. Retrieved from http://articles.latimes.com/2004/aug/22/opinion/oe-fuller22

Garcia J. 1995. *In Dallas: District Ranks School by How Much They Improve on Traditional Tests*. Catalyst (November). 1-8.

Garmannslund P, Elstad E, Langfeldt G. 2008. *Lærernes opplevelse av måling og rangering av kvalitetsaspekter ved undervisning og læringsprosesser*. In Langfeldt G, Elstad E and Hopmann ST (eds.), Ansvarlighet i skolen: Politiske spørsmål og pedagogiske svar. Oslo: Cappelen forlag.

Gavrielatos A. 2009 (November 23). *League tables don't tell a school's whole story*. Sydney Morning Herald.

Goddard M, Mannion R. 2004. *The role of horizontal and vertical approaches to performance measurement and improvement in the UK Public Sector*. Public Performance & Management Review. 28(1): 75-95.

Goldstein H, Spiegelhalter DJ. 1996. *League tables and their limitations: statistical issues in comparisons of institutional performance*. Journal of the Royal Statistical Society: Series A (Statistics in Society). 159: 385-443.

Goslin DA. 1963. *The Search for Ability: Standardized Testing in Social Perspective*. NY: Russell Sage.

Goslin DA, Epstein RR, Hallock BA. 1965. *The Use of Standardized Tests in Elementary Schools*. New York: Russell Sage.

Govindarajan V, Fisher J. 1990. *Strategy, Control Systems, and Resource Sharing: Effects on Business-Unit Performance*. The Academy of Management Journal. 33(2): 259-285.

Greener I. 2005. *The potential of path dependence in political studies*. Politics. 25(1): 62–72.

Gubman EL. 1998. *The Talent Solution: Aligning Strategy and People to Achieve Extraordinary Results*. McGraw Hill, New York.

Guha R, Adelman N, Arshan N, Bland J, Caspary K, Padilla C, Biscocho F. 2014. *Taking stock of the California Linked Learning District Initiative: Fourth-year evaluation report*. Menlo Park, CA: SRI International.

Haertel EH, Herman JL. 2005. *A historical perspective on validity arguments for accountability testing*. Yearbook of the National Society for the Study of Education. 104(2): 1-34.

Hamilton LS, Stecher BM, Klein SP. National Science Foundation (U.S.) (eds.). 2002. *Making sense of test-based accountability in education*. Santa Monica, CA: Rand.

Han Chun Y, Rainey H. 2006. *Consequences of goal ambiguity in public organizations.* Public Service Performance: Perspectives on Measurement and Management. 92-112.

Haney W. 1981. *Validity, vaudeville, and values: A short history of social concerns over standardized testing.* American Psychologist. 36(10): 1021–1034.

Hanushek EA, Benson CS, Freeman RB, Jamison DT, Levin HM, Maynard RA, Murnane RJ, Rivkin SG, Sabot RH, Solmon LC, Summers AA, Welch F, Wolfe BL. 1994. *Making Schools Work: Improving Performance and Controlling Costs*. The Brookings Institution, Washington DC.

Hanushek EA, Raymond ME. 2005. *Does school accountability lead to improved student performance?* Journal of Policy Analysis and Management. 24: 297–327.

Harel I, Papert S (eds.). 1991. *Constructionism.* Norwood, NJ.

Harlen W. 2007. *Criteria for Evaluating Systems for Student Assessment.* Studies in Educational Evaluation. 33(1): 15-28.

Heinrich CJ. 2002. *Outcomes–based performance management in the public sector: Implications for government accountability and effectiveness*. Public Administration Review. 62(6): 712–725.

Heinrich CJ. 2007. *Evidence-based policy and performance management challenges and prospects in two parallel movements*. The American Review of Public Administration. 37(3): 255–277.

Heinrich CJ, Marschke G. 2010. *Incentives and their dynamics in public sector performance management systems*. Journal of Policy Analysis and Management. 29: 183-208.

Hendrickson KA. 2012. *Assessment in Finland: A Scholarly Reflection on One Country's Use of Formative, Summative, and Evaluative Practices.* Mid-Western Educational Researcher. 25 (1/2): 33-43.

Heneman HG. 1999. *Assessment of the motivational reactions of teachers to a school-based performance award program*. Journal of Personnel Evaluation in Education. 12: 143-159.

Her Majesty's Inspectorate of Constabulary. 2000. *On the Record: Thematic Inspection Report on Police Crime Recording*. London, Home Office.

Herrera Santana HJ. 2007. *Problemas críticos de Colombia [Critical problems of Colombia]*. Bogota, Colombia: National university of Colombia.

Hershberg T. 2002. Comment. In D. Ravitch (Ed.), *Brookings papers on education policy*: 2002 (pp. 324-333). Washington, DC: Brookings Institution Press.

Heubert J, Hauser R (eds.). 1999. *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Heyman J, Ariely D. 2004. *Effort for payment: A tale of two markets*. Psychological Science. 15: 787-793. doi:10.1111/j.0956-7976.2004.00757x

Hodgson L, Farrell CM, Connolly M. 2007. *Improving UK public services: a review of the evidence*. Public Administration. 85(2): 355-382.

Holmes CT. 2006. *Low test scores + high retention rates = more dropouts*. Kappa Delta Pi Record. 42(2): 56–58.

Hood C. 2006. *Gaming in target world: The targets approach to managing British public services*. Public Administration Review. 66(4): 515–521.

Hopmann ST. 2008. *No child, no school, no state left behind: schooling in the age of accountability*. Journal of Curriculum Studies. 40(4): 417-456.

ICFES. 2010, May 30. *Estructura general del examen SABER 3,5,9 [SABER test: General structure for 3rd, 5th and 9th grades]*. Bogotá, Colombia. Retrieved from: http://www2.icfes.gov.co/estudiantes-y-padres/pruebas-saber-3-5-y-9-estudiantes/estructura-general-del-Examen

ICFES. 2017a. *SABER 3: Guía de orientación [SABER test for 3rd grade: Guidelines]*. Bogotá, Colombia.

ICFES. 2017b. *SABER 11: Guía de orientación [SABER test for 11th grade: Guidelines]*. Bogotá, Colombia.

Ittner CD, Larcker DF, Randall T. 2003. *Performance Implications of Strategic Performance Measurement in Financial Services Firms*. Accounting, Organisations and Society. 28(7/8): 715-741.

Jackson CK, Bruegmann E. 2009. *Teaching students and teaching each other: The importance of peer learning for teachers*. American Economic Journal: Applied Economics. 1(4): 85-108. doi:10.1257/app.1.4.85

Jacob BA. 2002. *Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools*. (Working paper 8968). Cambridge, MA: National Bureau of Economic Research.

Jacob BA. 2005. *Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools*. Journal of Public Economics. 89: 761–796.

Jacob BA, Levitt SD. 2003. *Rotten apples, an investigation of the prevalence and predictors of teacher cheating*. The Quarterly Journal of Economics. 118: 843–877.

Jacobs R, Martin S, Goddard M, Gravelle H, Smith P. 2006. *Exploring the determinants of NHS performance ratings: lessons for performance assessment systems*.

Jaeger RM. 1982. *The final hurdle: Minimum competency achievement testing*. In G. R. Austin and H. Garber (eds.). The Riseand Fall of National Test Scores. New York: Academic Press.

Jaworski BJ, Young SM. 1992. *Dysfunctional Behavior and Management Control: An Empirical Study of Marketing Managers*. Accounting, Organizations and Society. 17(1): 17-35.

Jensen MC. 2003. *Paying People to lie: the truth about Budgeting process*. European Financial Management. 9(3): 379-406.

Jones K, Tymms P, Kemethofer D, O'Hara J, Mcnamara G, Huber S, Myrberg E, Skedsmo G, Greger D. 2017. *The unintended consequences of school inspection: the prevalence of inspection side-effects in Austria, the Czech Republic, England, Ireland, the Netherlands, Sweden, and Switzerland*. Oxford Review of Education. 1-18.

Jordan PC. 1986. *Effects of an Extrinsic Reward on Intrinsic Motivation: A Field Experiment*. The Academy of Management Journal. 29(2): 405-412.

Kane T, Staiger D. 2002. *Improving school accountability measures*. (Working paper). Los Angeles, CA: University of California Los Angeles.

Kaplan RS, Norton DP. 1992. *The Balanced Scorecard Measures That Drive Performance*. Harvard Business Review. 70(1): 9.

Kaplan RS, Norton DP. 1993. *The balanced scorecard-Measures That Drive Performance; Putting the balance scorecard to work*. Harvard business review.

Kaplan RS, Norton DP. 1996. *The Balanced Scorecard:Translating Strategy into Action*. Harvard Business School Press, Boston.

Kellaghan T, Greaney V, Murray S. 2009. *Using the Results of a National Assessment of Educational Achievement*. Washington D.C.: The World Bank.

Kelley C. 1998. *The Kentucky School-Based performance Award Program: School-Level Effects*. Educational Policy. 12: 305-324.

Kelley C. 1999. *The motivational impact of school-based performance awards*. Journal of Personnel Evaluation in Education. 12: 309-326.

Kelley C, Protsik J. 1997. *Risk and reward: perspectives on the implementation of Kentucky's school-based performance award program*. Educational Administration Quarterly. 33: 474-505.

Kelman S, Friedman JN. 2009. *Performance improvement and performance dysfunction: An empirical examination of distortionary impacts of the emergency room wait-time target in the English national health service*. Journal of Public Administration Research and Theory. 19(4): 917–946.

Kemmerer FN, Windham DM (eds). 1997. *Incentives analysis and individual decision making in the planning of education*. UNESCO, International Institute for Educational Planning, Paris. 94-107.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. 2000. *What do test scores in Texas tell us?* Santa Monica, CA: Rand.

Klerks M. 2013. *The effect of school inspections: A systematic review*. Manuscript submitted for publication.

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

Kohn A. 1986. *No contest: The case against competition*. Boston, MA: Houghton Mifflin.

Kohn A. 1999. *Punished by Rewards: The trouble with gold stars, incentive plans, A's, praise and other bribes*. Houghton Muffin, New York.

Koretz D. 1996. *Using student assessments for educational accountability*. In Hanushek EA and Jorgenson DW eds. Improving America's Schools: The Role of Incentives. National Academy Press, Washington DC. 171-96.

Korte G. 2015. *The Every Student Succeeds Act vs. No Child Left Behind: What's changed?* USA Today news, published on December 11, 2015.

Kreps DM. 1997. *Intrinsic Motivation and Extrinsic Incentives*. The American Economic Review. 87(2): 359-364.

Kuhlmann D. 2012. *The Interaction of Contract, Control, and Relational Norms as Governance Mechanisms in IS Outsourcing Relationships.* Editor: Diplomarbeiten Agentur.

Kunz AH, Pfaff D. 2002. *Agency theory, performance evaluation, and the hypothetical construct of intrinsic motivation*. Accounting, Organization and Society. 27(3): 275-296.

Kupiainen S, Hautamaki J, Karjalainen T. 2009. *The Finnish education system and PISA.* Finland: Ministry of Education Publications.

Ladd H. 2007. *Holding Schools Accountable Revisited*. Spencer Foundation Lecture in Education Policy and Management, presented at the 2007 APPAM Fall Research Conference, Washington DC, November 8.

Ladd HF. 1999. *The Dallas school accountability and incentive program: an evaluation of its impacts on student outcomes*. Economics of Education Review. 18: 1-16.

Lan Z, Rainey HG. 1992. *Goals, Rules, and Effectiveness in Public, Private and Hybrid Organizations: More Evidence on Frequent Assertions About Differences*. Journal of Public Administration Research and Theory. 2: 5-28.

Langer EJ, Roth J. 1975. *Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task*. Journal of Personality and Social Psychology. 32: 951–955.

Lawler EE, Rhode JG. 1976. *Information and Control in Organizations*. Pacific Palisades: Goodyear Publishing Company.

Levine CH, Peters BG, Thompson FJ. 1990. *Public Administration: Challenges, Choices, Consequences.* Boston: Scott, Foresman/Little, Brown.

Levine D. 1994. *Instructional approaches and interventions that can improve the performance of African American students.* The Journal of Negro Education. 63(1): 46-58.

Levine FM, Broderick JE. 1983. *Attribution and contrast: Two explanations for the effect of external rewards on intrinsic motivation*. British Journal of Psychology. 74: 461-466.

Levitt R, Janta B, Wegrich K. 2008. *Accountability for Teachers: A Literature Review*. Santa Monica, CA: Rand Cooperation.

Li A, Maani K. 2011. *Dynamic Decision-Making, Learning and Mental Models.*

Lilliard D, DeCicca P. 2001. *Higher standards, more dropouts? Evidence within and across time*. Economics of Education Review. 20(5): 459–473.

Lingle JH, Schiemann WA. 1996. *From Balanced Scorecard to Strategic Gauges: Is Measurement Worth it?* American Management Association.

Linn RL. 1998. *Assessments and accountability*. CSE Technical Report 490. Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing.

Linn RL. 2000. *Assessments and accountability*. Educational Researcher. 29(2): 4-16.

Lipsky M. 1980. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. New York: Russell Sage Foundation, 1980. Politics & Society. 10(1): 116–116.

Llinás R. 1995. *Ciencia, educación y desarrollo: Colombia en el siglo XXI [Science, education and development: Colombia in the 21st century]*. Bogota, Colombia: Magisterium of education. 24-49.

Luginbuhl R, Webbink D, De Wolf I. 2009. *Do school inspections improve primary school performance?* Educational Evaluation and Policy Analysis. 31: 221–237.

Luna-Reyes LF, Andersen DL. 2003. *Collecting and analyzing qualitative data for system dynamics: methods and models*. System Dynamics Review. 19: 271-296.

Lynch RL, Cross KF. 1991. *Measure Up: Yardsticks for Continuous improvement*. Blackwell, Cambridge.

Lynn LE. 1996. *Public Management as Art, Science, and Profession*. Chatham NJ: Chatham House.

Maciejczyk A. 2016. *Challenges of Control in Functional Organization Structures: Example of Outsourcing Sector*. Journal of Economics and Management. 25: 49-62.

Marquès Graells P. 2001. *La enseñanza. Buenas prácticas. La motivación [The teaching. Good practices. The motivation]*. Barcelona, Spain: Autonomous university of Barcelona.

Mars G. 1982. *Cheats at Work – An Anthropology of Workplace Crime*. Allen & Unwin: London.

Martínez Rizo F. 2010. *Uso de SABER: Llamados de atención y lecciones [Use of SABER: calls of attention and lessons]*. Colombian ministry of education. Accessed October 20, 2018.

Marton F, Säljö R. 1976a. *On the qualitative difference in learning I-Outcome and Process.* British Journal of Educational Psychology. 46: 4-11.

Marton F, Säljö R. 1976b. *On the qualitative difference in learning II-Outcome as a function of the Learner's conception of the task.* British Journal of Educational Psychology. 46: 115-127.

MassPartners for Public Schools. 2005. *Facing reality: What happens when good schools are labeled "failures"? Projecting adequate yearly progress in Massachusetts schools.* Massachusetts Teachers Association.

Mathis WJ, Trujillo TM. 2016. *Lessons from NCLB for the Every Student Succeeds Act*. National Education Policy Center.

McCartney S, Brown R. 1999. *Managing by numbers: using outcome measures in the NHS*. International Journal of Health Care Quality Assurance. 12(1): 6-12.

McDonnell L. 2005. *Assessment and accountability from the policymaker's perspective*. In E.H. Haertel, J.L. Herman. (eds.). Uses and misuses of data for educational accountability and improvement: The 104th yearbook of the national Society for the Study of Education, Part II. Malden, MA: Blackwell

McGinnes S, Elandy K. 2012. *Unintended Behavioural Consequences of Publishing Performance Data: Is More Always Better?* The Journal of Community Informatics. 8(2).

Meier KJ. 1985. *Regulation: Politics, Economics and Bureaucracy*. New York: St. Martin's Press.

Mestres i Salud L. 2004. *Comunicación y pedagogía: Nuevas tecnologías y recursos didácticos [New technologies and teaching resources]*. Madrid, Spain: Center of communication and pedagogy. 37-39.

Meyer MW, Gupta V. 1994. *The performance paradox*. Research in Organizational Behaviour. 16: 309–369.

Micheli P, Neely A. 2010. *Performance Measurement in the Public Sector in England: Searching for the Golden Thread*. Public Administration Review. 70(4): 591-600.

Mintzberg H. 1979. *The structuring of organizations*. Englewood Cliffs, N.J., Prentice Hall.

Morecroft J. 2007. *Strategic modeling and business dynamics.* Chichester: Wiley.

National Governors' Association. 1989. *Results in Education: 1989*. Washington, D.C.

National Research Council. 2012. *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: The National Academies Press.

Neely A, Adams C, Crowe P. 2001. *The performance prism in practice*. Measuring Business Excellence. 5(2): 6-13. https://doi.org/10.1108/13683040110385142

Neely A, Kennerley M, Martínez V. 2004. *Does the balanced scorecard work: an empirical investigation*. Performance Measurement and Management: Public and Private. Performance Measurement Association, Edinburgh.

Nichols-Barrer I, Haimson J. 2013. *Impacts of five Expeditionary Learning middle schools on academic achievement*. Cambridge, MA: Mathematica Policy Research.

Nichols SL, Glass GV, Berliner DC. 2006. *High-stakes testing and student achievement: Does accountability pressure increase student learning?* Educational Policy Analysis Archives. 14 (1). http://epaa.asu.edu/epaa/v14n1/

Niemi H, Toom A, Kallioniemi A. 2016. *Miracle of Education: The Principles and Practices of Teaching and Learning in Finnish Schools*. Second Revised Education ed. Sense Publishers, Rotterdam.

Noble AJ, Smith ML. 1994. *Old and new beliefs about measurement-driven reform: "The more things change, the more they stay the same"* (CSE Technical Report No. 373). Los Angeles: University of California, Los Angeles, CRESST.

O'Day J. 2002. *Complexity, Accountability, and School Improvement.* Harvard Educational Review. 72(3).

Odden A, Kelley C. 1997. *Paying Teachers for what they know and do: new and smarter compensation strategies for Teachers*. Sage, London.

OECD. 1999. *Measuring Student Knowledge and Skills. A New Framework for Assessment*. OECD Publications, Paris.

OECD. 2004. *Learning for Tomorrow's World: First Results from PISA 2003*. OECD Publications, Paris.

OECD. 2005. *Formative Assessment: Improving Learning in Secondary Classrooms*. Paris: OECD Publishing.

OECD. 2010. *DAC Guidelines and Reference Series. Quality Standards for Development Evaluation*. Paris, France.

Oppel R, Shear M. 2014, May 28. *Severe report finds V.A. hid waiting lists at hospitals*. New York Times.

O'Reilly CA, Roberts KH. 1974. *Information Filtration in Organizations: Three Experiments*. Organizational Behavior and Human Performance. 253-265.

Ouchi WG. 1979. *A Conceptual Framework for the Design of Organizational Control Mechanisms.* Management Science. 25: 833-848.

Owens SA, Ranick DL. 1977. *The Greensville program: A commonsense approach to basics*. Phi Delta Kappan. 58(7): 531–533.

Parker W, Mosborg S, Bransford J, Vye N, Wilkerson J, Abbott R. 2011. *Rethinking advanced high school coursework: Tackling the depth/breadth tension in the AP US Government and Politics course*. Journal of Curriculum Studies. 43(4): 533-559.

Pedulla J, Abrams L, Madaus G, Russell M, Ramos M, Miao J. 2003. *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers [Electronic version]*. Boston: National Board on Educational. Testing and Public Policy, Lynch School of Education, Boston College.

Perry JL, Porter LW. 1982. *Factors Affecting the Context for Motivation in Public Organizations*. Academy of Management Review. 7: 89-98.

Perryman, J. 2007. *Inspection and emotion*. Cambridge Journal of Education. 37: 173–190.

Peterson's. 2010. *Peterson's Official Guide to Mastering Dsst Exams.*

Pfeffer J, Sutton RI. 2000. *The knowing-doing gap: How smart companies turn knowledge into action*. Boston, MA: Harvard Business School Press.

Pidd M. 2005. *Understanding perverse effects of public sector performance measurement*. Lancaster University Management School Working Paper.

Pil FK, Leana CR. 2009. *Applying organizational research to public school reform: The effects of teacher human and social capital on student performance*. Academy of Management Journal. 52: 1101-1124.

Platts KW, Sobotka M. 2010. *When the uncountable counts: an alternative to monitoring employee performance*. Business Horizon. 14(1): 28-42.

The constitution of Colombia. 1991. Bogota, Colombia: Legis.

Pollitt C. 2005. *Indicators are Indicators, Work is work and politics are politics: Do we have theories to handle this?* Erasmus University Rotterdam.

Popham WJ. 1991. *Why Standardized Tests Don't Measure Educational Quality*. Educational Leadership. 56(6): 8-15.

Popham WJ. 2001. *The truth about testing: An educator's call to action*. Hawaii, United States of America: Association for supervision and curriculum development.

Popham WJ. 2003. *Test Better, Teach Better: The Instructional Role of Assessment*. Alexandria, VA: Association for Supervision and Curriculums Development.

Popham WJ, Cruse KL, Rankin SC, Sandifer PD, Williams PL. May 1985. *Measurement-driven instruction: It's on the road*. Phi Delta Kappan. 66(9): 628–634.

Radnor Z. 2005. *Developing a typology of organisational gaming*. Presentation at the EGPA conference, Bern.

Rainey HG. 1993. *Toward a Theory of Goal Ambiguity in Public Organizations*. In J.L. Perry (ed.), Research in Public Administration, Vol. 2. Greenwich, CT: JAI Press.

Read WH. 1962. *Upward Communication in Industrial Hierarchies*. Human Relations. 3-15.

Resnick D. 1982. *History of educational testing*. In A. K. Wigdor and W. R. Garner (eds.). Ability Testing: Uses, Consequences, and Controversies, Part II. Washington, D.C.: National Academy Press, 173–194.

Rickles J, Zeiser KL, Yang R, O'Day J, Garet MS. 2019. *Promoting Deeper Learning in High School: Evidence of Opportunities and Outcomes.* Educational Evaluation and Policy Analysis. 41(2): 214–234. https://doi.org/10.3102/0162373719837949

Ridgway V. F. 1956. *Dysfunctional Consequences of Performance Measurements. Administrative Science Quarterly*. 1(2): 240-247.

Rinne R, Kivirauma J, Simola H. 2002. *Shoots of revisionist education policy or just slow readjustment? The Finnish case of educational reconstruction*. Journal of Educational Policy. 17: 643-658.

Roderick M, Bryk A, Jacob B, Easton J, Allensworth E. 1999. *Ending social promotion: Results from the first two years*. Chicago: Consortium on Chicago School Research.

Roeber E. February 10, 1988. *A history of large-scale testing activities at the state level*. Paper presented at the Indiana Governor's Symposium on ISTEP. Madison, Ind.

Ronen J, Sadan S. 1981. *Smoothing Income Numbers – Objectives, Means and Implications.*

Rosenthal L. 2004. *Do school inspections improve school quality?* Ofsted inspections and school examination results in the UK. Economics of Education Review. 23: 143–151.

Ross SM, Sanders WL, Wright SP, Stringfield S, Wang LW, Alberg M. 2001. *Two- and Three-Year Achievement Results from the Memphis Restructuring Initiative.* School Effectiveness and School Improvement. 12(3): 323-346.

Rothstein R. 2008. *Holding accountability to account: How scholarship and experience in other fields inform exploration of performance incentives in education*. National Center on Performance Incentives Working paper 2008-04, Vanderbilt University, Peabody College.

Rumberger RW, Larson KA. 1998. *Student mobility and the increased risk of high school dropout*. American Journal of Education. 107(1): 1–35.

Rustique-Forrester E. 2005. *Accountability and the pressures to exclude: A cautionary tale from Accountability Texas-Style England*. Education Policy Analysis Archives.

Ryan JE. 2004. *The perverse incentives of the No Child Left Behind Act*. New York University Law Review. 79: 932-989. doi: 10.2139/ssrn.476463

Salazar Rua RS. 2016. *A Dynamic Performance Management approach to study the impact of education actors' decision making on the Colombian public schools' educational quality.* European Master Thesis in System Dynamics.

Santos SP, Belton V, Howick S. 2002. *Adding value to performance measurement by using system dynamics and multicriteria analysis*. International Journal of Operations & Production Management. 22(11): 1246.

Saunders M, Lewis P. 2012. *Doing reresearch in Business and Management. An essential guide to planning your project*. Harlow, England: Financial Times Prentice Hall.

Schiller K, Muller C. 2000. *External examinations and accountability, educational expectations, and high school graduation*. American Journal of Education. 108(2): 73–102.

Schleicher A. 2007. *Can competencies assessed by PISA be considered the fundamental school knowledge 15-year-olds should possess?* Journal of Educational Change. 8: 349–357.

Schmelkes S. 1995. *Hacia una mejor calidad de nuestras escuelas. Secretaría de Educación Pública. [Towars a better quality of our schools]*. Mexico City, Mexico: Secretary of public education.

Scogin SC, Kruger CJ, Jekkals RE, Steinfeldt C. 2017. *Learning by experience in a standardized testing culture: Investigation of a middle school experiential learning program*. Journal of Experiential Education. 40(1): 39-57.

Scriven M. 1996. *Types of evaluation and types of evaluator*. American journal of evaluation. 17: 151-161.

Shadow Strategic Rail Authority (SSRA). 2000. *On Track: A new way of measuring performance.* London, OPRAF.

Shalley CE, Oldham GR, Porac JF. 1987. *Effects of Goal Difficulty and Expected External Evaluation on Intrinsic Motivation*. The Academy of Management Journal. 30(3): 553-563.

Shaw I, Newton DP, Aitkin M, Darnell R. 2003. *Do Ofsted inspections of secondary schools make a difference to GCSE results?* British Educational Research Journal. 29: 63–75.

Shohamy E. 2001. *The power of tests: A critical perspective on the uses of language tests*. Harlow, UK: Pearson Education.

Siciliano MD. 2007. *The Effect of School Choice on Segregation and Inequality in Charlotte Mecklenburg Public Schools: An Analytic Framework Utilizing Spatial and Probability Models.* Association of Public Policy and Management Conference. Washington, DC

Siciliano MD. 2017. *Professional Networks and Street-Level Performance: How Public School Teachers' Advice Networks Influence Student Performance*. The American Review of Public Administration. 47(1): 79–101.

Sills D. 1957. *The Volunteers: means and Ends in a National Organization*. Glencoe, IL: Free Press.

Simon DR, Eitzen DS. 1986. *Elite Deviance*. Allyn & Bacon, Boston.

Simons R. 1995. *Levers of Control: How Managers Use Innovative Control Systems to Drive Strategic Renewal*. Harvard Business School Press, Boston.

Sipkoff M. 2007. *Go Carefully When Measuring Quality*. Managed Care. 16(9/September 2007): 25-26, 28, 31.

Skinner BF. 1938. *The Behavior of Organisms: An Experimental Analysis*. New York: D. Appleton-Century Company, Inc.

Skinner W. 1969. *Manufacturing-missing Link in Corporate Strategy*. Harvard Business Review.

Smith F. 1986. *High school admission and the improvement of schooling*. New York: New York City Board of Education.

Smith KB. 1994. *Politics, Markets, and School Bureaucracy: Reexamining School Choice.* Journal of Politics. 56: 475-91.

Smith KB. 1998. *What Price Commonwealth: Democracy, Markets, and Education*. Mimeo. University of Nebraska.

Smith ML, Fey P. 2000. *Validity and accountability in high-stakes testing*. Journal of Teacher Education. 51(5): 334-344.

Smith P. 1990. *The Use of Performance Indicators in the Public Sector*. Journal of the Royal Statistical Society, Series A. 153: 53-72.

Smith P. 1993. *Outcome-related Performance Indicators and Organizational Control in the Public Sector*. British Journal of Management. 4: 131-151.

Smith P. 1995. *On the Unintended Consequences of Publishing Performance Data in the Public Sector*. International Journal of Public Administration. 18(2/3): 277-310.

Soobaroyen T. 2005. *Management control systems and dysfunctional behaviour: an empirical investigation.*

Sterman JD. 1994. *Learning in and About Complex Systems*. System Dynamics Review. 10(2-3): 291–330.

Stiefel L, Schwartz AE, Rubenstein R, Zabel J. (eds). 2005. *Measuring School Performance and Efficiency: Implications for Practice and Research*. American Education Finance Association 2005 Yearbook. Larchmont, NY: Eye on Education.

Strand S. 1997. *Key Performance Indicators for Primary School Improvement*. Educational Management Administration & Leadership. 25(2): 145-153.

Strobel J, Van Barneveld A. 2009. *When is PBL More Effective? A Meta-synthesis of Meta-analyses Comparing PBL to Conventional Classrooms*. Interdisciplinary Journal of Problem-Based Learning.

Sturman, L. 2003. *Teaching to the test: Science or intuition?* Educational Research. 45: 261–273.

Terman JN, Yang K. 2016. *Reconsidering Gaming in an Accountability Relationship: The Case of Minority Purchasing in Florida*. Public Performance & Management Review. 40(2): 281-309.

Thomas JW. 2000. *A Review of Research on Project Based Learning*. The Autodesk Foundation. Project Based Learning (PBL). California, USA.

Tighe S. 2019. *Rethinking Strategy: How to anticipate the future, slow down change, and improve decision making*, 1st Edition.

Todd DP. 2000. *A "Dynamic" balanced socrecard*. Management Science and Information Systems. Auckland, University of Auckland.

Torres J, Duque H. 1994. *El proceso de descentralización educativa en Colombia [The decentralization process of the education in Colombia*. Colombian review of the education.

Tyack DB. 1974. *The one best system: A history of American urban education*. Cambridge, MA: Harvard University Press.

Tyack DB, Cuban L. 1995. *Tinkering Toward Utopia: A Century of Public School Reform*. Cambridge: Harvard University Press.

Tymms, P. 2004. *Are standards rising in English primary schools?* British Educational Research Journal. 30: 477–494.

UNESCO (The United Nations Educational, Scientific and cultural organization). 2005. *Promover la seguridad humana: Marcos éticos, normativos y educacionales en América Latina y el Caribe [Promoting human security: Ethical, normative and educational frameworks in Latin America and the Caribbean]*. Paris, France.

Vaishnav A. 2005. *Adding Value to Student Assessment: Does "Value-Added Assessment" Live up to its Name?* Harvard Education Letter. 21(3): 1-3.

Van Thiel S, Leeuw FL. 2002. *The performance paradox in the public sector*. Public Performance & Management Review. 267-281.

Vaughn D. 1983. *Controlling Unlawful Organizational Behaviour – Social Structure and Corporate Misconduct*. University of Chicago Press, Chicago.

Vasquez Heilig J, Darling-Hammond L. 2008. *Accountability Texas-Style: The Progress and Learning of Urban Minority Students in a High-Stakes Testing Context*. Educational Evaluation and Policy Analysis. 30(2): 75–110. https://doi.org/10.3102/0162373708317689

Visscher AJ, Coe R. 2003. *School performance feedback systems: conceptualisation, analysis, and reflection*. School effectiveness and school improvement. 14(3): 321-349.

Walker A, Leary H. 2009. *A Problem Based Learning Meta-Analysis: Differences Across Problem Types, Implementation Types, Disciplines, and Assessment Levels.* Interdisciplinary Journal of Problem-Based Learning.

Warren K. 2008. *Strategic Management Dynamics*. John Wiley & Sons: Chirchester, UK.

Weibel A, Rost K, Osterloh M. 2010. *Pay for performance in the public sector—Benefits and (hidden) costs*. Journal of Public Administration Research and Theory. 20: 387-412. doi:10.1093/jopart/mup009

Weiner MJ. 1980. *The effect of incentive and control over outcomes upon intrinsic motivation and performance*. Journal of Social Psychology. 112(2): 347-354.

Werner RM, Asch DA. 2005. *The unintended consequences of publicly reporting quality information*. JAMA: the journal of the American Medical Association. 293(10): 1239.

Wheelock A. 2003. *School awards programs and accountability in Massachusetts: Misusing MCAS scores to assess school quality*. Cambridge, MA: Fair Test.

Wiggins A, Tymms P. 2002. *Dysfunctional effects of league tables: A comparison between English and Scottish primary schools*. Public Money and Management, 22: 43–48.

Wilcox B, Gray J. 1996. *Inspecting schools: Holding schools to account and helping schools to improve*. Buckingham: Open University Press.

Wilson JQ. 1989. *Bureaucracy: What Government Agencies Do and Why They Do It.* New York: Basic Books.

World Bank. 2006. *The World Bank Annual Report 2006*. Washington, DC: World Bank.

World Bank. 2017. *World Bank Annual Report 2017*. Washington, DC: World Bank.

Wouters M, Wilderom C. 2008. *Developing performance-measurement systems as enabling formalization: A longitudinal field study of a logistics department*. Accounting, Organizations and Society. 33(4): 488–516.

Yang T-M, Maxwell TA. 2011. *Information-sharing in public organizations: A literature review of interpersonal, intra-organizational and inter-organizational success factors*. Government Information Quarterly. 28: 164-175. Retrieved from http://dx.doi.org/10.1016/j.giq.2010.06.008

Youngblood AD. 2003. *The Use of Multi-Attribute Utility Theory to Address Trade-Offs for the Balanced Scorecard*. Department of Industrial Engineering Fayetteville, University of Arkansas. 222.

Zambrano MF. 2015. *Alcances e inconsistencias del índice sintético de calidad educativa, diseñado y aplicado por el ministerio de educación de Colombia en el año 2015 [Scope and inconsistencies of the synthetic index of educational quality, designed and applied by the Ministry of Education of Colombia in 2015].* Bogota, Colombia: IQ Matrix foundation.

Zhang Y, Shaw JD. 2012. *Publishing in AMJ – part 5: Crafting the methods and results*. Academy of Management Journal. 551: 8-12.

Zucker S. 2004. *Administration Practices for Standardized Assessments*. Pearson Assessment Report.