School governance through performance-based accountability: A comparative analysis of its side effects across different regulatory regimes

Introduction

Over the past two decades, performance-based accountability (PBA) has gained increasing popularity in school governance, endorsed for its potential to enhance the effectiveness and equity of educational systems. The United States presents a key example of PBA's growth in school governance, primarily through national standardized tests. With the implementation of the No Child Left Behind Act (NCLB) in 2001, the United States exemplifis the growth of PBA in school governance, primarily through annual nationwide standardized tests[1]. This approach was further solidified with the introduction of the Race To The Top Act in 2009 and Every Student Succeeds Act in 2015 (Amrein-Beardsley & Holloway, 2017; Baker et al., 2013; Portz & Beauchamp, 2022). Beyond the United States, in recent decades, the adoption of PBA policies has experienced exponential growth globally (Verger et al., 2019b). That is, PBA instruments have been enacted in countries with diverse institutional, economic and regulatory characteristics, such as England, Norway, Germany, the Netherlands, Italy, Chile, Brazil and South Africa, among others (Högberg & Lindgren, 2021; Lingard et al., 2015; Spreen, 2004). PBA motivates school actors to align their objectives and practices with the learning standards and achievement goals outlined in assessment frameworks. Confronted with regulatory pressures, schools are expected to embark on organizational improvement initiatives, involving professional development, strategic planning, and seeking external support. The underlying assumption is that if external pressures align effectively with internal improvement processes, there will be a corresponding increase in school achievement (Polikoff, 2012). PBA systems, designed to enhance system-level efficiency, aim to ensure that school practices align with

[1] NCLB was a landmark U.S. federal law aimed at improving K-12 education. It emphasized standards-based education reform, holding schools accountable for student performance through standardized testing. NCLB sought to close achievement gaps, increase teacher quality, and ensure that all students, regardless of socioeconomic background, received a quality education.

administrative goals. Their objective is to foster more effective teaching methods and promote educational equity, aiming to ensure a consistent level of proficiency in core subjects for all students, irrespective of their backgrounds (Lingard et al., 2017).

However, numerous studies indicate that the enactment of PBA policies is not always straightforward and does not uniformly generate the anticipated responses within schools (Hofflinger & von Hippel, 2020; Mintrop, 2004; Lowenhaupt et al., 2016). A growing body of research has identified side effects of PBA, which imply unintended effects that arise in addition to the policy's primary objectives, across various educational domains (Au, 2007; Berryhill et al., 2009). These side effects have raised concerns among researchers and policymakers, particularly because they disproportionately impact the educational experiences of disadvantaged groups of students (Özek, 2015).

Despite the steady growth in research that focuses on PBA side effects, the literature remains fragmented and characterized by significant gaps. In particular, a systematic analysis of the mechanisms that trigger different side effects under specific conditions is still lacking, with the role of policy design features in influencing the emergence of these side effects being particularly poorly understood. That is, the absence of comprehensive studies that analyze the emergence of side effects across different accountability regimes, and especially between both high- and low-stakes regimes, impedes a full understanding of how policy settings and organizational characteristics influence these outcomes. Consequently, theories that explain the emergence of PBA side effects in different policy contexts are not yet identifiable.

To address some of the gaps outlined so fat, in this chapter, we systematically review empirical studies where the identification of PBA side effects is a central focus. Our configurative systematic literature review (SLR) brings together a total of 133 studies that have analyzed this phenomenon globally. Our main aims are to: 1) determine under what conditions PBA side effects are more likely to occur and the causal mechanisms triggering their emergence; 2)

enhance the understanding of how educational contexts, policy design features and the perceptions of local actors contribute to these side effects; and 3) identify research gaps and propose future lines of research.

The paper is organized as follows. After this introduction, we delve into the concept of side effects from a social science perspective, with a specific focus on their application in the realm of performance indicators and evaluation. The following section provides a foundational understanding of the PBA policy program, outlining its key features, policy design characteristics, and the theoretical underpinnings expected to drive specific outcomes. Next, we detail our review methodology, describing the processes of data collection and analysis. The next two sections are dedicated to presenting our core findings, categorizing side effects into various domains (professional, pedagogical, organizational, student-related, and external contexts) and elucidating the underlying mechanisms. The final section synthesizes our main conclusions.

The intended, unintended and side effects of public policy: the case of performance targets and quantification

The concept of side effects, which gained prominence in the mid-20th century through philosophers like Karl Popper (1944) and Jean-Paul Sartre (1960), was significantly integrated into the social sciences, largely due to Robert Merton's seminal 1936 work on 'The Unanticipated Consequences of Purposive Social Action' (Vernon, 1979). Merton's analysis became instrumental in understanding the often paradoxical interplay between intentionality (marked by self-reflection and reflexivity) and the unexpected benefits and/or adverse outcomes resulting from social action (Baert, 1991). Rather than being an exception, to Merton and his followers, unintended consequences are the pervasive, recurrent logical result of factors

such as incomplete information, imperfect enactment, and unforeseen interpersonal interactions prevalent in all types of complex social systems.

By emphasizing the interconnectedness and complexity of social environments, Merton's (1936) work echoes systems theory, according to which side effects are emergent properties of a larger system that goes through change. Alterations in one part of the system can have ripple effects throughout the system, resulting in unanticipated consequences (Meadows et al., 1982). This overarching principle is applied across various research domains, notably in policy transfer studies, which highlight the numerous side effects arising from borrowing policies between contexts without sufficient consideration of contextual differences, as discussed by Dolowitz & Marsh (2000).

The concept of side effects underwent systematic development within functionalist sociology, and subsequently expanded to a wider spectrum of theories relevant to policy studies (Baert, 1991). For instance, according to rational choice theory, side effects may emerge as a result of individuals strategically pursuing their self-interests in response to new policy environments. When faced with new regulations, social actors strategically adjust their behavior to maximize personal benefits, potentially leading to individuals gaming the system. In a similar line of reasoning, principal-agent theory posits that side effects can arise when agents prioritize their interests (or preferences) over faithfully executing the principal's intentions. Agents may have more and better information than the principal and/or face conflicting goals, leading to deviations from the intended course in project implementation (Müller & Turner, 2005). From a complementary perspective, behavioral scientists consider that actors might game the system if they are pushed to attain goals perceived as too difficult or when the consequence at stake is perceived as high (Patrick et al., 2018).

More recently, the sociology of quantification has directed specific attention toward performance indicators and evaluation as factors that motivate behavioral change and

potentially lead to unintended consequences. This body of literature conceives individuals as reflexive and adaptive actors who react to social measures, often doing so in an amplified manner. Espeland and Sauder (2007) contributed to this line of inquiry by identifying specific mechanisms that drive people's (hyper-)reactivity to numerical data. The first mechanism, derived from Merton (1948), involves self-fulfilling prophecies. These are processes where "reactions to social measures confirm the expectations or predictions that are embedded in measures or which increase the validity of the measure by encouraging behavior that conforms to it" (Espeland & Sauder, p.11). The second mechanism, commensuration, entails the comparison of diverse entities using a common metric, influencing behavior by affecting "what we pay attention to, which things are connected to other things, and how we express sameness and difference" (Espeland & Sauder, 2007, p.16).

These mechanisms possess constitutive effects, creating new relationships, meanings, and practices that may result in both homogenization and standardization, and attempts to manipulate the system as shortcuts to meet specific targets. Nonetheless, it is important to note that side effects do not uniformly manifest everywhere. Their intensity varies based on factors such as the type of incentives attached to performance achievement, professional power and influence, organizational procedures, and managerial styles. These factors help to understand the variegated constitutive effects of numbers (Dahler-Larsen 2014).

The analysis of PBA serves to illustrate the complex interplay between the explicit objectives of accountability policies and the unforeseen benefits or adverse outcomes that emerge from social action. Specifically, this case offers an opportunity to discern the extent to which the constitutive effects of numerical data are contingent upon contextual factors. PBA represents an instance of a policy intervention heavily dependent on performance indicators and targets, the side effects of which have been extensively reported but not systematically compared across diverse regulatory regimes. In this chapter, we undertake an examination of

how PBA policy manifests in varied contexts, aiming to unravel how policy design and institutional characteristics, among other factors, play a role in influencing actors' behavior within the educational domain.

PBA: policy features and intentions

Accountability is not a new concept in the governance of education. Historically, education systems have implemented forms of bureaucratic and professional accountability (Darling-Hammond, 2004). However, over the last 2 decades, PBA has become increasingly central. This form of accountability focuses on goal-setting and performance achievement, specifically targeting students' learning outcomes as the primary educational quality indicator. It often uses growth measures or the achievement of minimum thresholds[2] and involves creating data infrastructures that cover students' and schools' performance along with contextual data, as measured through large-scale assessments (LSAs).[3] Teachers and school leaders are the prime account-givers concerning the academic performance of pupils, whereas the main forum is the public administration, but not only since other educational stakeholders can also interact as account-takers (Hooge et al., 2012).

Nonetheless, far from homogeneous, PBA systems can follow very different approaches (Högberg & Lindgren, 2021). The academic literature has usually distinguished high- from low-stakes accountability approaches, based on the type of consequences, incentives and uses of assessment results, and the pressure the system is expected to place on school actors (Au, 2007; Maroy & Pons, 2019).

High-stakes accountability regimes attach significant consequences to assessment results. They rely heavily on material consequences, including financial incentives for schools

[2] Growth measures imply students are evaluated on the basis of how much they have progressed academically over time. Minimum thresholds involve setting a minimum standard or benchmark that students are expected to reach, thereby serving to determine whether students have achieved a predetermined level of proficiency.

[3] Large-scale assessments imply the systematic evaluation of the skills or knowledge of a significant number of students within a specific education system or across different systems. Large-scale assessments are often conducted on a regional, national or international level.

and/or school actors, promotions or demotions for teachers and principals, and the occasional closing of underperforming schools. High-stakes accountability regimes can also be related to students' graduation decisions, thereby also contributing to pressure on students. They often involve the publication of test results, which adds reputational pressure on schools, and in contexts of free school choice, can alter the demand pool of schools. This accountability regime tends to conceive school actors as benefit-maximizing individuals who will put in more effort to promote students' learning if the right structure of incentives is in place (e.g., Ehren et al., 2015) .

Low-stakes accountability regimes involve assessments with minimal consequences, primarily aimed at providing feedback and support to schools for improvement purposes. Although these consequences are not punitive, they may still result in increased oversight for underperforming schools. This approach assumes that providing information about student learning -including their achievements in different knowledge domains, types of competences that students struggle the most with, the identification of underachieving groups, and so on- will encourage schools to design more accurate improvement strategies, leading to gradual improvement over time. The consequences of the system are more ideational and symbolic rather than material, with increasing transparency in the educational system being another implicit goal of PBA. In the case of low-stakes accountability, individuals are viewed as social actors who aim to comply with social expectations and norms, but also as refelctive actors who will respond to relevant data and information, especially when these data highlight areas for improvement and growth (Maroy & Voisin 2013).

Despite these differences between high- and low-stakes systems, the two approaches expect to generate similar positive impacts in at least four ways. First, the assessment of performance metrics will encourage teachers and school leaders to focus and devote more effort to raising student performance expectations and improving their learning (Dee & Jacob, 2011).

Second, test results will serve to identify good practices and develop evidence-based actions, which can inspire others, while discarding efforts that do not necessarily contribute to improvement (Matteucci et al., 2017). Thirdly, the data generated by the external tests will be of great value to identify groups of students or schools with greater difficulties and, subsequently, to adopt more focused compensatory measures (Chiang, 2009). Fourth, PBA will help to create a culture of 'continuous improvement' and transparency in schools and even foster a sense of shared responsibility and collaboration among teachers for student performance (Ehren et al., 2015).

Nonetheless, PBA policies often deviate from their anticipated outcomes, resulting in a diverse array of side effects. Our review seeks to explore whether these side effects are not only prevalent in accountability systems with high-performance pressure but also occur in low-stakes systems with more symbolic types of consequences.

Methodology

Recent education research invites us to see side effects as inseparable from intended effects, and to aim to anticipate the side effects of any intervention, as experimental sciences do in fields such as medicine (Zhao, 2017). Because most purposes, rationales, and motivations are known retrospectively, previous research can contribute to overcoming the epistemological difficulties in anticipating side effects. An SLR of a policy intervention whose side effects are widely acknowledged but not always systematically mapped can be a fine strategy to advance this research agenda. Methodologically, this research is based on a configurative SLR (Gough et al., 2012; Petticrew & Roberts, 2006). To carry out our SLR, we followed the steps indicated in the PRISMA protocol for developing SLRs (Moher et al., 2012). The main phases for conducting our SLR are presented below (see also Figure 1).

Identification

Our search strategy was mainly based on electronic searches in two international databases: SCOPUS and Web of Science (WoS). These are considered two of the most reliable databases in social sciences research (Harzing & Alakangas, 2016; Zhu & Liu, 2020). Research on PBA policies is cross-disciplinary. In our review, we focused on investigations published in fieldss such as social sciences, arts and humanities, psychology, and economics. We included empirical studies published between 2000-2022 that focused on compulsory education. Although we did not restrict the geographical scope of our review, we did limit the search based on the publication language, selecting only research published in English, Spanish, Italian, French, Portuguese, Afrikaans and Dutch. Additionally, we used a filter for the type of publication, including only journal articles, books and book chapters.

Given our focus on PBA policies in education, we used keywords such as "accountability", "new public management", "school-based management", "benchmarks", "testing", "league table", "ranking", "high-stakes" and "low-stakes" to find relevant literature related to this topic. We used different terms to capture the multiple instruments and policy design options of PBA (Levatino et al., 2024; Maroy & Voisin, 2017). Given our focus on compulsory education, we selected the terms "school," "teacher," "school principal," “school leader” and "head teacher" to identify relevant literature on school responses to PBA policies. Finally, given our focus on policy side effects, we used the terms "unexpected," "undesired," "unintended", and "side effects." These are the most common terms by which the existing literature refers to the side effects of PBA policies (Falabella, 2020; Thiel & Bellmann, 2017; Zhao, 2017). After conducting the initial search, we obtained a sample of 1,205 articles. After applying the filters, this number was reduced to 439 articles (see Figure 1).

Screening

Following SLR protocols (Pawson et al., 2005), we carried out two rounds of screening in which six researchers participated. In the first round of screening, each piece of research was reviewed by a minimum of two authors. In this phase, the reviewers read the titles and abstracts and verified that the papers met the inclusion criteria. After the first round of screening, we eliminated a total of 299 articles (we excluded 88 duplicate documents and 211 documents that did not meet some of the inclusion criteria mentioned above).

Eligibility and inclusion

After the first round of screening, we generated a database of 140 documents. In the second screening, we followed the same process based on the aforementioned criteria, but unlike the first stage, we analyzed the full texts. In this phase, we excluded a total of 25 articles. The primary reasons for exclusion were that upon reviewing the content, it was evident that the publications were not empirical investigations, did not primarily focus on PBA policies, or failed to analyze the enactment and side effects of PBA instruments. In addition, during the analysis of the articles, the authors identified key references that had not appeared through the search strategy in the aforementioned databases. Following this process, 19 publications that met the inclusion criteria of the review were incorporated. After performing the three steps mentioned above, we obtained a final dataset consisting of a total of 133 documents.

(FIGURE 1 ABOUT HERE)

Data analysis

We followed a realist synthesis approach - the most typical configurative review approach - to understand why, how and under what circumstances side effects of PBA policies emerge (Pawson et al., 2005; Greaves et al., 2023; Gough et al., 2012). With this purpose in mind, we elaborated a codebook that can be consulted in the Appendix (see Table 2), on which the qualitative content analysis of the included documents was based. For the elaboration of the codebook, we included deductive (theory-driven) and inductive codes. After carefully reading the abstract of all papers and an exploratory coding of 12 articles, we identified five distinct domains of side effects: professional, pedagogical, organizational, student and external contexts. To organize our analysis, we categorized the research by country and by predominant accountability models in each context (high- and low-stakes accountability). The distinction between high- and low- stakes PBA regimes follows the frameworks established by Euryidice (2020) and Högberg & Lindgren (2021).

In total, we assigned around 20 to 30 articles to analyze to each team member. Each article had a person in charge of coding. The team members conducted a detailed reading of each text assigned to them, assessing the quality and validity of each investigation. To strengthen the validity of the SLR we followed the PRISMA protocol and guidelines as detailed by Page et al. (2021). In addition, to guarantee interreliability of the coding process, we followed a two-step strategy. First, we organized four coding retreats (four to five hours each) to code simultaneously on the ATLAS.ti web. All team members participated in these coding retreats and discussed any doubts and difficulties that arose during the coding process to try to harmonize the criteria. Second, the whole team coded a sample of 5 papers selected randomly to guarantee the consistency of the coding process. In addition, we used an online forum to share our analytic doubts during the coding process.

Mapping the side effects of PBA

We organize the diverse array of identified side effects into five distinct domains: professional, pedagogical, organizational, students and external contexts. However, it is important to note that this categorization, although aiding in data organization and analysis, is not entirely definitive. The porosity of this classification stems from two main factors: a single side effect may overlap across multiple categories, and most side effects often arise from a confluence of various causal factors. This often results in a 'cascade' effect, where interconnected consequences manifest in several domains simultaneously.

Professional

In the professional domain, PBA is expected to strengthen deliberation about instruction among teachers, inform professional development policy, and make school professionals more committed to improving the learning outcomes of their students. However, many of the identified side effects undermine the essence of teacher professionalism, including autonomy and discretion, work collaboration and ethical practice. .

Teacher well-being and workforce challenges

Although some teachers gain satisfaction and a sense of self-efficacy from high scores (Holloway & Brass, 2018), PBA pressure predominantly affects teachers' well-being negatively. Numerous articles see external evaluation and accountability instruments as a source of anxiety, frustration, and stress that demoralizes teachers and affect their connection to the profession. In some cases, this tendency manifests in an increase in sick leaves and even in an increase in attrition rates (Neal, 2010; Penninckx, et al., 2016). Teachers' well-being issues are particularly well-documented in the United States (Haney, 2000; Nichols & Berliner 2005, 2007), where blaming dynamics and fears of losing jobs intensified since the enactment of NCLB in 2002 (Vasquez Heilig, 2011). In most states, schools' scores are published, and students' results are attributed to teachers' abilities.

Reputational concerns and the desire to be seen as a 'good teacher' explain to some extent the state of anxiety (Booher-Jennings, 2005). In some cases, it appears that teachers are stressed because they do not know how to manage what they perceive as an inherent tension between performance and students' inclusion goals within the same accountability mandate (Russell & Bray, 2013). Nonetheless, research also shows that school leaders can play an important role in mediating pressure and the emotional reactions of teachers (Jaffe-Walter, 2023).

In Europe, together with the school principal (Bukh et al., 2022; Camphuijsen, 2021), the inspection service tends to be the agent passing or modulating the pressure that teachers experience. Jones et al. (2017), who compare the effects of PBA in a sample of European countries that range from high- to low-stakes, showed that professional effects such as an administrative burden and paperwork are rather related to inspection styles instead of accountability pressure, with some inspectors being more hands-on and performance-data-oriented than others.

These studies show that high levels of performance pressure are not only present in high-stakes regimes. Pressure is relative and subjectively experienced. Teachers may experience high pressure if they were previously in a low-pressure environment. In Belgium, for example, even though the educational system is considered low-stakes, it still causes significant stress and anxiety among school staff (Penninckx et al. (2016). This is because in a context with little emphasis on evaluation and transparency, any form of assessment is viewed as having substantial consequences. As Penninckx et al. (2016, p.16) put it, "the limited experience of Flemish schools with accountability measures [...] makes the schools perceive the stakes to be higher than they actually are." Literature produced in continental Europe frequently identifies PBA as a source of administrative burden, which is often linked to increased job dissatisfaction and eventually brings teachers and, mainly, school leaders, to

"fabricate" documentation merely to please the inspection service, without this documentation having any practical implication at the school level (Penninckx, 2016; Pagès, 2021).

Teacher attrition, a severe indicator of teacher dissatisfaction, is notably reported as a direct consequence of PBA pressure in the US and England. In these countries, the challenging work environment created by accountability demands is making it increasingly difficult to attract and retain high-quality teaching staff. Attrition worsens in situations where teacher evaluation is associated with external assessments and growth and value-added models[4]. Teachers often view these models as unfair and dismissive of their efforts, particularly when working with disadvantaged students (Hewitt, 2015). The increased workload and pressure are more acute in underperforming schools, leading to greater staffing challenges (Ladd, 2002). In such disadvantaged schools, the tendency to blame teachers for poor results heightens their inclination to leave, resulting in more frequent teacher shortages and turnover compared to other schools (Neal, 2010)

Teacher attrition is also influenced by individual characteristics. Boyd et al. (2008) found that more experienced teachers often leave their jobs rather than adapt their teaching styles to meet accountability standards. Shirrell (2018) analyzed the impact of race, discovering that attrition rates are lower among Black teachers working with underperforming Black students, suggesting a role for racial solidarity in mitigating this effect. It is also influenced by the design of PBA policies. For example, Abrams et al. (2003) compared states with high-stakes and low-stakes testing in the US, concluding that the intensity of accountability pressures significantly affects teacher dissatisfaction and attrition. Additionally, the impact of

[4] Growth models and value-added models are statistical approaches used in education to measure the progress or improvement of students over time. Growth models focus on tracking the academic progress of individual students over a specific period. Value-added models aim to quantify the contribution of teachers, schools, or educational interventions to students' academic growth by accounting for the background characteristics of students. The idea is to measure the "value" that a teacher or school adds to students' learning beyond what would be expected. In the academic literature, concerns have been expressed about the accuracy of both measurements (e.g. Amrein-Beardsley, 2008).

PBA extends beyond teachers in tested subjects; instructors in areas such as drama, music, dance, and performing arts are increasingly facing job scarcity due to a reduced emphasis on non-tested subjects (Gewirtz et al., 2021).

Teacher autonomy and school governance

In many countries, teaching is considered a semi-profession because professional standards are largely defined by external authorities rather than exclusively within the profession itself (Demirkasımoğlu, 2010). PBA has exacerbated this trend, leading to increased influence of psychometricians, test specialists, and public management experts on the content and methods of teaching in schools. Indeed, numerous studies show that due to the pressures of PBA, teachers frequently experience diminished professional autonomy and control.

Accountability pressures are often seen as undermining teachers' creativity and spontaneity in the classroom (Perryman et al., 2011, Wills & Sandtholz, 2009). Berkovich (2019) shows that the erosion of teachers' autonomy has increased over time in the United States and in Australia, in parallel to an intensification of performance pressure. Kaynak Elcan (2020) confirms this trend with a study conducted in the US Midwest, in which all teachers interviewed "shared the perception that they had less power and less control over what they were doing in the classroom now than in the past, which reduced the pleasure they took in teaching" (p. 30). Even in schools whose strong point is an alternative pedagogic approach (Scott, 2017) or in which school leaders are supportive of teachers (Will & Sandholtz, 2009), teachers experience that their pedagogic autonomy is constrained because of PBA.

Alongside these developments, many teachers report feeling that their capacity for professional judgment has been compromised. They often struggle to adhere to their preferred teaching styles or feel constrained in delivering content they believe would be most beneficial

for their students (Macqueen et al., 2018; Russell & Bray, 2013; Thiel, 2021). However, this shift in judgment capacity is not always portrayed negatively in the academic literature. Datnow and Park (2018) suggest that because teachers' judgments can be biased, the challenge to professional judgment posed by accountability might on occasion be beneficial. “Performance data”, they argued, “can play a very powerful role in challenging stereotypes and providing an opportunity for educators to examine the relationship between instructional practices and achievement” (p. 146). Furthermore, Penninckx et al. (2016) contended that, under certain circumstances, PBA can actually enhance teachers' sense of self-efficacy.

Other studies indicate that PBA significantly alters the internal relationships and hierarchical structures within schools. The dynamics of horizontal relationships among school staff are often disrupted due to the accountability placed on school leaders by educational authorities (Fitzgerald, 2009). These dynamics create a principal-agent relationship where school leaders, facing pressure from superintendents or inspection services, may pass this pressure onto their teachers (Penninckx et al., 2016). However, the effects of these dynamics vary based on the existing culture and leadership styles within each school (Buisson-Fenet & Pons, 2019). In fact, not all outcomes are negative; in some instances, properly managed external pressure can lead to more constructive feedback from principals to teachers (Donaldson & Woulfin, 2018).

Several studies highlight that PBA also erodes collegiality among teachers. As collaborative efforts diminish, competitive dynamics emerge. This is evident when teachers responsible for tested grades attribute poor student performance to their lower-grade colleagues (Booher-Jennings, 2005), when there are continuous comparisons of teachers' results, and when teachers in tested subjects or departments monopolize resources (Perryman et al., 2011). The competitive atmosphere is particularly pronounced in environments where market forces and accountability demands intersect, as observed in a comparative study between England and

Scotland (Wiggins & Tymms, 2002), or in systems where teachers' salaries are influenced by value-added measures (Hewitt, 2015).

Nonetheless, other studies acknowledge that, by promoting competition between schools and defining performance goals, PBA can have the effect of strengthening teacher cohesion (Luna & Turner, 2001). Oyarzún Vargas and Falabella (2020, p. 12-13) observe that, in the Chilean context, accountability pressure makes teachers feel "motivated and committed to maximizing school performance and also, very importantly, to share responsibility, blame, and feelings of guilt". To some observers, however, the first cycles of PBA generate a sense of enthusiasm and positive stress, but once they are routinized, accountability and assessments lose such a motivational capacity (Penninckx et al., 2016).

Professional Standards and Ethical Challenges

In response to the pressures of testing, some teachers may deviate from professional standards, including, for instance, attempts to manipulate test results by changing the composition of the testing group or assisting certain students during tests (Amrein-Beardsley et al., 2010; Ohemeng & McCall-Thomas, 2013). Although these practices are predominantly observed in high-stakes environments like certain states in the United States and Canada (Collins, 2014; Rezai-Rashti, 2020), research indicates that such tactics can be triggered not only by material consequences but also by symbolic and reputational pressures (Levatino et al., 2024).

Pedagogical

In the pedagogical domain, PBA is expected to support schools in identifying areas of improvement and professional learning and to help teachers in identifying students in need of assistance. Although some of the reviewed articless indeed find that PBA policies can support

school actors in doing so (e.g. Polesel et al., 2014), various pedagogical side effects are simultaneously identified.

Teaching to the test and curriculum narrowing

The most commonly identified side effects in the pedagogical domain are teaching to the test and curriculum narrowing. Teaching to the test typically entails teachers adapting their instruction to closely align with the content and format of the standardized test, thereby aiming to raise test performance (e.g. Amrein & Berliner, 2002; Diamond, 2007; Jacob, 2005; Ehren & Hatch, 2013; Haney, 2000; Nichols & Berliner, 2005, 2007). The reviewed articles highlight how teaching to the test can take different forms. In many cases, teachers engage in emphasizing the specific topics or sets of skills that the standardized test will address (e.g. Abrams et al., 2003), which sometimes involves practices of drilling and rote memorization (e.g. Polesel et al. 2014). teaching to the test can also involve a focus on strengthening students' test-taking abilities. For example, Abrams et al. (2003) highlight how teachers in different states in the US engage in practices such as demonstrating how to mark answer sheets correctly, teaching test-taking skills, and providing students with test-taking tips. To do so, some teachers (in particular those working in high-stakes accountability states) rely on specific test preparation materials developed commercially or by the state (Abrams et al., 2003). Reliance on practice tests as well as simulation of the testing situation during a regular school day are also commonly reported on in the reviewed literature (e.g. Diamond, 2007). Finally, Ehren and Hatch (2013) showed that teachers can also align their own classroom and formative assessments to the standardized test (in terms of content, format or rubrics) in order to familiarize students with the test format.

Teaching to the test differs from sporadic test preparation. Whereas preparing students for the test is a punctual action before an upcoming assessment, oriented to ensure students

know what to expect (to manage potential anxiety) and become familiar with the test format (e.g. understand how to answer multiple-choice questions), teaching to the test is a more prolonged strategy aimed at securing good results on the test. An important distinction between the two practices lies in their intent. Familarizing students with the test format ensures the test accurately reflects their abilities without necessarily interfering with their understanding of the subject matter. On the contrary, teaching to the test involves a narrow focus on test-specific content and risks compromising broader educational goals by prioritizing short-term gains in test performance. This practice may also inflate student results, undermining the validity of standardized tests as accurate measures of broader learning outcomes.

In the case of curriculum narrowing, the instructional time that is spent on knowledge areas that are tested is increased, at the expense of areas/subjects that are not, such as social studies, science, physical education, technology or music (e.g. Polesel, Rice & Dulfer, 2014; Wills et al., 2009; Meadows, 2018). In only one of the reviewed studies, conducted in England, a minority of surveyed teachers (4%) hinted at the opposite practice: curriculum expansion. They agreed with the statement that the reforms would ensure "a broader and more balanced curriculum than before" (Gewirtz et al., 2021, p. 517).

The majority of articles explain the practices of teaching to the test and curriculum narrowing by pointing towards the excessive pressures that teachers feel to obtain high test performance. This is often the result of the consequences that teachers in some contexts face, including financial incentives, staff replacements or the threat of school closure. For this reason, it is perhaps not surprising that the majority of studies that reported on teaching to the test and/or curriculum narrowing were conducted in high-stakes accountability contexts such as England and many states in the United States (e.g. Amrein-Beardsley & Berliner, 2002; Diamond, 2007; Jacob, 2005; Nichols & Berliner, 2005, 2007).

However, both side effects are also found in other contexts, including low-stake regimes where material consequences attached to test performance are minimal or absent such as Italy (Landri, 2021), Israel (Feniger et al., 2016), and Germany (Thiel et al., 2017). To understand the emergence of teaching to the test and curriculum narrowing in low-stakes contexts, scholars have offered different explanations. For example, Thiel and Bellman (2017) argue that although the degree to which side effects occur may be influenced by features of the accountability system such as the stakes attached, the existence of side effects across accountability regimes may be understood as "systematic effects of accountability in education". That is to say, side effects might occur almost by default with the implementation of any type of accountability in education. Other scholars have pointed towards the power of numbers and evaluations as a central explanation for the emergence of side effects in low-stakes contexts. Feninger et al. (2015), for example, argued that "the use of external standardized tests, in itself, causes a shift in the way actors in the educational field think and speak about education" (p.3).

An often proposed solution to reduce the risk of side effects in pedagogical practice has been to rely on multiple measures to define what counts as success and to hold educational actors accountable accordingly. In the case of multimeasure accountability systems, test scores are, for example, complemented with district-wide or city-wide inspections or quality reviews of schools. Nonetheless, Ehren and Hatch (2013) show that side effects such as teaching to the test and curriculum narrowing also occur in multimeasure systems.

Teaching strategies

Although scholars have highlighted how PBA policies have a stronger impact on instructional content compared to pedagogy (Diamond, 2007), the reviewed articles show that some not only change what they teach (i.e. the knowledge and skills that they emphasize), but also their teaching methods and the ways they engage students with the instructional content . Various

studies, for example, identify teachers that turn to lecture-based, teacher-centered pedagogies in order to cover the material for the standardized test (e.g. Gewirtz et al., 2021). These pedagogical changes often go at the expense of more interactive forms of instruction and innovation (Hargreaves, 2020). Other studies document teachers who engage in instruction that focuses on recitation, memorization and low-level skills, instead of meaningful learning (Polesel et al., 2014). As such, these studies highlight how increased emphasis is placed on giving information rather than on creating a learning environment in which students can formulate questions, participate in meaning-making, and solve problems. Moreover, various studies report an increase in standardized instruction based on externally determined goals and activities (e.g. Russel & Bray, 2013; Wills & Sandholtz, 2009). Only one of the reviewed articles (Cuban, 2007) documented an expansion of student-centered pedagogies.

At the same time, not all schools and all students are equally affected by the occurrence of pedagogical side effects. According to Garner et al. (2017), low-performing schools, which often serve students from historically marginalized backgrounds, are disproportionately affected, thereby contributing to a paradoxical effect: A measure intended to address educational inequalities, such as accountability, ends up exacerbating them. As they argued, "teachers of students who underperform on standardized achievement tests are incentivized to `reteach` instead of being incentivized to teach for deeper understanding" (p. 421). These instructional strategies are particularly pronounced in high-stakes systems, such as the United States, England and Canada. Nonetheless, studies focused on countries such as the Netherlands and Sweden also hint at side effects related to how teachers teach. For example, Ehren et al. (2015) compared the impact of different inspection systems in European countries on the actions of school-level stakeholders. The analysis showed that school leaders working in countries where the Educational Inspectorate relied on a differentiated inspection model, such

as in the Netherlands, Sweden and England, were more likely to report that they felt the need to discourage teachers from experimenting with new teaching methods.

Organizational

In the organizational domain, PBA encourages school actors to rearrange their internal setups and allocate resources more effectively and efficiently to improve students' learning, especially in basic skills and subjects that are tested regularly. Our review highlights that in systems with high-stakes accountability, schools often direct their resources (including extra teaching time) to subjects and/or departments that are tested. This shift often goes at the expense of subjects and/or departments that are not part of the performance measurement and thus are seen as 'less crucial' (Booher-Jennings, 2005; Ehren & Hatch, 2013; Jacob, 2005; Wills & Sandholtz, 2009; Wiggins & Tymms, 2002). Several reviewed studies reveal that this situation often leads to increased competition or rivalry between departments (e.g. Perryman, 2011).

In the Unites Sates, particularly, there is ample evidence of 'strategic staffing'. Teachers who are considered 'little effective' are moved to teach subjects and/or courses that are not tested, while 'highly effective' teachers are assigned to tested subjects and/or grade levels (Feniger et al., 2015; Henry et al., 2022; Thiel et al., 2017). Another example of strategic staffing comes from Denmark, where financial incentives were introduced for underperforming schools in order to motivate them to improve student performance. Bukh et al. (2022) showed that following the introduction of the financial incentive, school leaders began to overspend on hiring additional staff. This overspending, thereby exceeding the allocated budget, was aimed at meeting the school performance targets (Bukh et al., 2022).

Another organizational side effect that we identified by means of our review is the practice of altering the test group (or 'reshaping the test pool'), which implies that low-performing students are excluded from taking the test (Ehren & Hatch, 2013; Nichols & Berliner 2005, 2007; Meadows & Black, 2018). This practice is mainly identified in high-

stakes accountability regimes and involves different strategic behaviors. For instance, several studies reveal that educators label students as 'special education' or 'disabled' to avoid their participations in the tests (Booher-Jennings 2005; DeMatthews & Knight, 2019; Ehren & Swanborn, 2012; Fetler, 2019; Figlio & Getzler, 2006; Jacob, 2005; Shirrell, 2016). Other studies indicate that certain students are held back in the grade before the test is administered (Ehren & Swanborn, 2012; Rezai-Rashti & Segeren, 2020). Evidence furthermore reports the practice of cream-skimming, which involves pushing students 'out of school' or not enrolling low-performing students. These kinds of behaviors occur in both high-stakes accountability and low-stakes accountability systems (e.g. Thiel & Bellmann, 2017).

In addition to practices aimed at reshaping the test pool, various studies provide evidence of educational triage, where students are sorted based on test scores, leading to a heavy focus on borderline students, also called 'bubble-kids' (Bukh et al., 2022; Collins, 2014; Datnow & Park, 2018; Diamond & Cooper 2007; Jacob, 2005; Hargreaves, 2020; Perryman et al., 2011; Vasquez Heilig et al., 2012; Wilson et al., 2004; Wiggins & Tymms, 2002). These practices often involve ability grouping, where students are placed in either advanced or less advanced curricular pathways and groups (Ehren & Hatch, 2013; Garner et al., 2017; Park & Datnow, 2017). These organizational arrangements may be seen as problematic by certain teachers, who notice a conflict between educational triage and a more student-centered, inclusive education approach (Horn, 2018).

Students

Side effects on students are explicitly reported only in articles referring to high-stakes PBA systems. According to the reviewed articles, these side effects often emerge as indirect consequences of the professional, pedagogical and organizational changes brought about by PBA which aim to boost test scores.

Management of student diversity

Part of the research reviewed shows that, in some cases, flexible learning groups resulting from PBA data analysis have been identified as a way to tailor teaching to students' needs, dedicating them extra time, attention and accommodations (Datnow & Park, 2018; Figlio & Gezler, 2006). Nonetheless, more often, practices like educational triage, ability grouping are related to fixed, rigid, reified views of their abilities, reinforcing ideas of inherent strengths or weaknesses. Besides, they often mean that effort, attention and resources are focused on specific groups of students (the so-called 'bubble students') while neglecting and limiting learning opportunities for others, usually, over- and underachievers, special needs' students, students with behavioral problems and s on. (Perryman et al., 2011; Bertrand & Marsh, 2021; Meadow & Black, 2018; Wiggins & Tymms, 2002; Collins, 2014; Wilson et al., 2004). As also highlighted in the subsection on the organizational domain, underachieving students might unsuitably be classified as having a disability (DeMatthews & Knight, 2019) and be placed into special education or retained in lower grades (Figlio & Gezler, 2006; Jacob, 2005). This misclassification can perpetuate negative beliefs, creating a dichotomy between 'smart' and 'dumb' students, thereby legitimizing deficit-thinking approaches.

The implicit or explicit labeling and categorization of students that accompany these practices affect for how students perceive themselves, impacting their self-esteem and, consequently, their educational chances (Horn, 2018). This phenomenon was called by Munoz-Chereau et al. (2022 , p. 15) the "fabrication of losers" with the consequent risk of the normalization of students' labeling carrying the longer-term risk of self-fulfilling prophecies (Horn, 2018). Labeling and student demotivation (Luna & Turner, 2001), together with test anxiety (Ohemeng & McCall-Thomas, 2013), and the feeling of having been neglected (Bianchi-Salazar, 2020), have been connected with increased likelihood to drop out and absenteeism (e.g. Luna & Turner, 2001). By analyzing administrative data from North Carolina, a high-staked PBA context where schools are exposed to being labeled as "failing"

and incurring sanctions, Holbein and Ladd (2017) found an increase in students' misbehavior among high- and low-performing students. The authors suggested misbehavior was linked to students' perception of them receiving less attention. In a study conducted in England, Rustique-Forrester (2005) found a climate of intolerance towards students with academic and behavioral difficulties because they were perceived as posing a threat to performance in school rankings and during inspections.

The reviewed research indicates that students with particular characteristics - such as race, social class, ability and primary language - face a higher likelihood of missing graduation (Kearns, 2011), being placed in special education (Figlio & Gezler, 2006), being placed in lower grades (Jennings & Beveridge, 2009), and experiencing psychological consequences (Whitney & Candelaria, 2017). This trend aggravates the marginalization of certain students, particularly those from lower socio-economic status and minority groups. It significantly affects equity by leading to less focus or even the marginalization of underperforming students from lower socio-economic backgrounds (Bianchi & Salazar, 2020; Booher-Jennings 2005; Holbein & Ladd 2017; Rustique-Forrester, 2005), ultimately resulting in 'a harmful policy for a school's weakest population' (Feniger et al., 2015, p. 13). In environments affected by systemic racism, it can further reinforce teachers' biases (Bertrand & Marsh, 2021; Horn, 2018).

Emotional effects

PBA side effects on students are also the result of the centrality of testing in accountability procedures. Students' side effects deriving directly from the tests are, for example, connected to 'peer competition or pressure' to complete the test before others do (Cho & Ebehard, 2013, p. 12), or to anxiety, fear of performance, and pressure to be examined, even in the absence of important consquences for them (Hargreaves, 2020). According to Gewirtz et al. (2021, p. 519), these emotional and psychological effects are connected to an increased focus on tests

and the ‘strengthening of an exam culture’. They seem to be especially intense, or at least more often reported on, in PBA contexts where material sanctions are at stake for schools and teachers (Cho & Ebehard, 2013; Muñoz-Chereau et al., 2022), or in contexts where free school choice policies are combined with the public dissemination of league tables (Gewirtz et al., 2021). Hargreaves (2020) suggests that the root cause lies in the preoccupation of children that they will be blamed by their teachers in case of bad test results. Student blaming, especially directed at low-performing ones, is also found by Vasquez Heilig et al. (2012) in Texas, where the stakes are very high for teachers and schools (replacement of staff, reduction of resources for schools, public embarrassment) and can be attributed to the pressure exerted by the district administration on school and teachers. In this paper, a school leader reported on the ethical dilemma he experienced when having to decide between potential school closure and “forcing” low-performing students “out of school” to improve results and avoid closure (p. 576). These dynamics have, in turn, an impact on trust relations between students and educators (Vasquez Heilig et al., 2012).

The erosion of trust between educators and students has also been reported as a consequence of teachers losing their emotional connection to the profession (Van Wyk & Le Grange, 2016) and their sensitivity to students’ needs (Hargreaves, 2020). A decrease of students' trust in teachers is also reported by Macqueen et al. (2018). In this case, however, it is attached to teachers feeling guilty “to impose an impossible task” on their students without being allowed to lend them a hand (p.13). Indeed, the study is about the psychological and academic difficulties students from Indigenous communities in Australia experience because of national standardized testing. These difficulties are related to the test language (English), which is not their mother tongue, and to the test content, exclusively based on Western urban values and knowledge, which is seen as neglecting Australian Indigenous communities’ oral background and culture. The same study reports how the negative effects of testing on these

students strongly depend on the emphasis schools and teachers place on tests and on the pressure they place on students. If this emphasis is accompanied by practices aimed at training students to the test, the authors also report the problem of providing this group of students a less meaningful learning experience.

All these results show how, in cases where PBA does not directly impact students, side effects on students are significantly influenced by the school environment and the degree to which their educators face pressure and accountability from various stakeholders. In contrast, in contexts with high-stakes for students (usually graduation), psychological effects come directly from the stake. As Kearns (2011) reported in the case of Ontario (Canada), some students are concerned about their future or experience a "shock" when they fail a test (p. 118). This feeling seems to originate from a discrepancy between their self-perception and their test performance, which undermines their self-esteem. Bad test results are indeed sometimes interpreted by students as signals of their failure and might generate "loser feelings" (Kearn, 2011, p. 119).

In addition to the impacts that predominantly affect specific student groups, the literature also reveals side effects that influence all students. Due to curriculum narrowing and extensive teaching-to-the-test activities, students receive less holistic (Collins, 2014) and less meaningful instruction (Bianchi & Salazar, 2020; Holbein & Ladd, 2017) Moreover, less time is spent on subjects like music, sport and technology, with potential consequences for how students approach learning (Hinnant-Crawford, 2019) and with potential long-term effects on employability and health (Wiggins & Tymms, 2002).

The impact on students, whether positive or negative, is influenced by the school's environment and how educators interpret and use standardized test data. According to Datnow and Park (2018), when the data are used to target instructional interventions on students perceived as crucial for improving scores, PBA can lead to adverse side effects. However, if

data are used to examine students' progress and weaknesses, and to tailor instruction to their needs, students can benefit significantly. This positive outcome is more likely in schools fostering cooperation and a shared sense of responsibility among teachers (Datnow & Park, 2018).

External context

External context side effects refer to the effects on the education market and school community, that is, effects on school composition, segmentation of educational supply, and changes in the perception of school reputation within the community, among others. As mentioned prevously, one of the main objectives of PBA policies is reducing educational inequalities. Nonetheless, paradoxically, existing research shows that a recurrent side effect of PBA is linked to increased school segmentation. This trend is observed in countries with a longer tradition of high-stakes accountability regimes, such as the United States, England, and Chile (Muñoz-Chereau et al. 2022; Davis et al., 2015).

Existing literature indicates that PBA can alter demand and supply behaviors in school-choice processes, which contribute to increasing school and urban segregation. From the demand side, McArthur and Reeves (2022) find that PBA policies affect families' residential mobility patterns in England. These authors argue that the publication of standardized test scores favors parental school choice and exacerbates processes of a geographic concentration of wealthy parents, which, in turn, may negatively affect the social and educational opportunities of children from disadvantaged backgrounds. Similarly, Basu (2004) finds that PBA policies contribute to labeling practices, categorizing schools based on performance, thereby exacerbating social polarization and inequalities between neighborhoods. This tendency can perpetuate and sharpen inequality between different urban areas or districts, further

deteriorating the reputation and attractiveness of disadvantaged neighborhoods (Bianchi & Salazar, 2022).

From the supply side, Ohemeng and McCall-Thomas (2013), in a qualitative study carried out in Ontario (Canada), found that the publication of standardized test results in the form of league tables or rankings by local think tanks may create additional performance pressure and promote competitive dynamics among schools. The publication of league tables can alter the boundaries of 'lived education markets' and scale up comparisons between schools (McArthur & Reeves, p. 518). Relatedly, recent research conducted in Canada suggests that "public shaming by the numbers contributes to stigmatization of schools in low-income, racially diverse neighborhoods where test results are often lower" (Rezai-Rashti & Segeren, 2020, p. 13; see also Muñoz-Chereau et al. 2022). PBA policies can also reinforce institutional hierarchies in local education markets (McArthur & Reeves, 2022). Research conducted in Australia also illustrates how PBA policies have altered schools' logics of action, increasing the school resources devoted to marketing activities, which are subsequently deviated from resources that could go to pedagogical improvement (Polesel et al. 2014). Similar effects on schools' logics of action have been documented in Canada and Hong Kong, where school leaders (especially those working in low-performing schools) increasingly invest time in managerial tasks to sell their school (Rezai-Rashti, 2020; Tse, 2018).

Finally, our review reveals that PBA instruments trigger changes in school culture across different accountability regimes. This shift is not only about enacting marketing activities but also involves practices that directly or indirectly lead to the exclusion of particular student groups (Darling-Hammond, 2004; Luna & Turner, 2001; Thiel & Bellmann, 2017; Vasquez Heilig et al., 2012). Nonetheless, alongside these adverse consequences, PBA policies can reshape power relations in school governance. By increasing transparency through the publication of test results, they may bolster parent empowerment and foster stronger teacher-

parent collaboration, although this effect is more pronounced in advantaged schools (Thiel, 2021).

Discussion: Comparing PBA side effects in high- and low-stakes regimes

Our study reveals side effects in both high- and low-stakes PBA regimes. The variance in types and frequencies of these side effects (see Table 1) is largely due to most PBA research focusing on high-stakes environments, particularly in countries like the United States, England, and Australia. These countries not only have a deeper history of educational research[5] but also a longer-standing tradition of PBA policies, allowing for more extensive study over time. Additionally, the prevailing assumption that higher stakes lead to more pronounced effects—both expected and unexpected—has likely rendered low-stakes settings less appealing to researchers. Nonetheless, albeit with less intensity and prevalence compared to high-stakes environments, side effects are reported in low-stakes PBA regimes as well. For instance, 'teaching to the test,' a widespread practice in the United States and England, also occurs in low-stakes countries like Germany, Belgium, and Italy, through with certain nuances. A notable difference is that in low-stakes contexts, specific side effects typically manifest in particular types of schools, whereas in high-stakes environments, these effects are more uniformly distributed across various school types.

Another difference shown in Table 1 is that research in low-stakes contexts tends to focus proportionally more on professional and external side effects. This trend may be related to the high value placed on professional autonomy and equitable forms of school provision in continental Europe. Side effects that impact these domains might receive more attention from researchers studying European educational systems, as they are perceived to be more disruptive.

[5] Of all the 1,826 most cited scholars in education all over the world (percentile 98 and above), 1,209 (two thirds of them) are based in these three countries (Ioannidis, 2023).

Table 1. Mapping PBA side effects in different regimes

Regime - countries most represented	**High-stakes** US, England, Canada (British Columbia, Ontario), Australia, Chile, NZ, Colombia, South Africa	**Low-stakes** Belgium, Denmark, Norway, France, Spain, Germany, Italy, Austria
Outcome dimension	**Side effects** (from more frequently reported to less)	
Pedagogic	● Teaching to the test (19) ● Curriculum narrowing (13) ● Teacher-centered pedagogies; traditional teaching (11) ● Adjusted perceptions of students as more or less capable (2) ● Flexible learning groups (2) ● Curriculum expansion (1) ● Student-centered pedagogies (1) Total: 49	● Teaching to the test (3) ● Curriculum narrowing (1) Total: 4

Students	● Marginalization/exclusion (22) ● Wellbeing (10) ● Students´ needs not prioritized (10) ● Less meaningful learning experience (8) ● Life chances (7) ● Dropout/absenteeism/ misbehavior (6) ● Less trust in teachers (5) ● Competition between students (1) Total: 69	● Marginalization/exclusion (1) Total: 1
Professional	● Wellbeing (14) ● Contradicting professional standards/Gaming (13) ● Individualism vs Cooperation (12) ● Autonomy (9) ● Attrition and shortages (7) ● Judgement (4) ● Hierarchies (3) ● Administrative burden (2) Total: 64	● Administrative burden (3) ● Wellbeing (2) ● Hierarchies (2) ● Judgement (2) ● Autonomy (1) ● Contradicting professional standards/gaming (1) Total: 11
Organizational	● Reshaping the test pool (21) ● Channeling resources to test subjects (14) ● Educational triage (14) ● Ability grouping (7) ● Conflicting inclusion goals (4) Total: 61	● Channeling resources to test subjects (3) ● Reshaping the test pool (2) ● Educational triage (1) Total: 6

External	● School segregation (6) ● Student exclusion (3) ● School quality (2) ● Intensification of school marketing (1) ● Hierarchies: schools' status (1) ● Misrepresentation or misuse of external tests data (1) ● Competition (1) Total: 15	● Intensification of school marketing (2) ● Misrepresentation or misuse of external tests data (2) ● School segregation (1) ● Student exclusion (1) ● Hierarchies: schools' status (1) ● Parents participation (1) Total: 8

Our review indicates that the mechanisms leading to side effects in PBA systems are influenced by the type of the regime. In high-stakes regimes, side effects frequently often arise from a fear of sanctions, which is logical given the explicit incentives and severe consequences often involved, such as job termination or school closure (Nichols & Berliner, 2005, 2007). The prospect of sanctions appears to be a primary catalyst for side effects. Similarly, the fear of losing market competitiveness is a significant concern in settings where poor performance may lead to a damaged reputation and a subsequent decline in parental preference, potentially diminishing the school's demand and corresponding resources (Rustique-Forrester, 2005).

Emerging evidence of side effects in low-stakes accountability systems, which often intentionally lack material consequences, prompts scholars to reassess underlying assumptions and seek new explanations and mechanisms. One such mechanism is commensuration, the practice of evaluating and comparing different entities using a common metric. This process has significant cognitive and political implications, influencing how educational professionals perceive and categorize students, schools, and their own work (Espeland & Sauder, 2007). It often leads to aligning notions of good teaching with performance metrics, creating self-induced pressure. Relatedly, commensuration can heighten reputational concerns regarding external audiences beyond educational authorities, such as families and the media. In contexts

where the media actively report on test results, their praise or criticism serves as a symbolic form of reward or punishment for educators and school administrators. This mechanism occurs irrespective of whether formal rewards or sanctions are imposed by education authorities (Ryan & Deci, 2020). In environments where reputational concerns are tied to perceptions of professional competence, anxiety about being blamed for not fulfilling public duties drives these concerns. This anxiety can be prompted by external pressures or self-imposed pressures, reflecting the profound impact of societal expectations on educational practices.

Another key social mechanism involves so-called 'instrument constituencies.' This term refers to groups of actors who emerge around policy instruments, such as large-scale assessments and performance metrics, as they develop, maintain, and promote these instruments (Béland & Howlett, 2016). Instrument constituencies can inadvertently turn policies into independent power sources, diverging from their original policy intentions. A notable example is seen in several European countries, where inspection services have evolved into such constituencies. Test results offer them objective and comparable performance metrics, greatly facilitating their oversight of schools. Consequently, through such interactions, inspectors exert more pressure on schools than initially planned, urging them to achieve or surpass certain performance benchmarks (Skedsmo et al., 2020).

The fact that mechanisms like instrument constituencies or reputational concerns are less commonly identified in high-stakes contexts does not imply their absence there. The comparison between high- and low-stakes regimes remains somewhat inconclusive. In low-stakes settings, it is unclear whether side effects would intensify if the stakes were raised. Conversely, in high-stakes systems, symbolic consequences often exist alongside material ones. Therefore, it is a simplification to attribute side effects in these contexts solely to the fear of sanctions. The other identified mechanisms likely interact in complex ways, potentially reinforcing one another.

Finally, our systematic review shows that the emergence and intensity of side effects in PBA regimes are influenced not only by the inherent nature of the regime (high- or low-stakes), but also by its interaction with various stakeholders (e.g. inspection, the media), school characteristics (leadership style, collegiality), approaches to inclusive education, and teachers' working conditions. For example, a school climate characterized by collegiality and supportive relationships can alleviate the fear of sanctions, whereas concerns about individual reputation may be heightened in settings with high teacher turnovers, precariousness and job insecurity. In contexts with low teachers' salaries, financial rewards attached to performance can significantly impact teachers' behavior. Media representation can furthermore have a demoralizing effect for school actors working in underperforming schools or in schools in a weaker position in the local education market.

Conclusions

In conducting a systematic review of the side effects of PBA across various educational contexts, we have identified a range of effects, spanning from well-documented to lesser-known or underexplored areas. Effects such as teaching to the test, curriculum narrowing, and altering the test pool are well-recognized in education research. However, subtler, long-term effects such as the erosion of teacher solidarity and diminishing student trust in educators, have received less attention.

Analytically, side effects crystallize in multiple domains: professional, pedagogical, organizational, students-related, and external. Nonetheless, most side effects are interconnected and emerge from a complex interplay of factors across these domains. For example, a decline in trust between teachers and students not only exacerbates a focus on testing in teaching (pedagogical), but can also prompt schools to alter their testing pools (organizational).

Similarly, pedagogical and organizational side effects affect not only teachers' work but also students' learning experiences. This is the case of practices like ability grouping that despite some praise for providing tailored education, often results in student labeling and categorization, significantly affecting students' self-esteem and educational opportunities. In fact, minority students, defined by race, social class, ability or mother tongue, are disproportionately impacted by PBA, even in the absence of high-stakes for students.

At the same time, although many side effects of PBA undermine its intended goals of enhancing efficiency, quality, and equity in education, we also note some positive outcomes, like increased family engagement in education, more focused professional development support and improved formative feedback from school leaders to teachers (Donaldson & Woulfin, 2018; Thiel, 2021; Wills & Sandholtz, 2009). The scarcity of identified positive side effects might stem from two factors. First, the terms 'side effects', 'unintended' or 'unexpected', that we used in our search (see Appendix Table 1), commonly implies negative connotations, influencing research results. Second, PBA's underlying assumptions overlook crucial educational dynamics, placing excessive emphasis on learning metrics and external incentives to shape behavior and drive commitment to school improvement. In essence, PBA policies are based on a reductive understanding of educational outcomes and a behaviorist view of human nature. This approach disregards the intrinsic, professional, and vocational motivations fundamental to education, often causing the actions of school actors to diverge from policy expectations.

At times, distinguishing between desired and undesired effects in the accountability debate relies heavily on subtle nuances and implicit assumptions. For example, what some researchers view as a beneficial effect of PBA, like increased performance pressure and competition among providers, others might see as detrimental. Similarly, determining the tipping point at which teaching to the test becomes harmful at the school level is also unclear.

Although many concur that a moderate level of test-focused teaching helps students to familiarize themselves with the test format and alleviate test-related anxiety, intense preparation throughout the year is widely recognized as harmful. This latter practice is considered not only degrading educational quality but also compromising the validity of large-scale assessments (Abrams et al., 2003). In a similar manner, focusing school resources on subjects central to assessment frameworks might undermine comprehensive, holistic, and inclusive educational approaches according to critics (e.g., Bianchi & Salazar, 2020), but emphasizing core subjects like Mathematics, Science, and Literacy is also an aspiration of many PBA advocates. These instances illustrate that the distinction between beneficial and adverse (side) effects is often subtle and nuanced. To a great extent, they highlight that the accountability debate is shaped by subjective educational and political notions of what constitutes quality in education.

Our research, encompassing studies from both high- and low-stakes PBA regimes, provides unique comparative insights. Recent literature has started to question the singular impact of accountability's material consequences on behavioral side effects. Our study builds on this trend by identifying and contrasting a wider array of social mechanisms that trigger side effects across different accountability contexts. Additionally, we identify which side effects and their underlying mechanisms are specific to particular contexts. In high-stakes environments, concerns such as potential sanctions and loss of market competitiveness are the most frequently documented, but we should not assume they are the sole or primary drivers of PBA side effects. In low-stakes contexts, side effects arise from factors such as commensuration, reputational concerns, and the influence of instrument constituencies. However, these factors can also impact high-stakes systems. Therefore, our research highlights the importance of examining how seemingly “soft mechanisms” influence school actors’ practices in both low- and high-stakes settings.

Finally, our findings encourage a closer look at factors that shape PBA processes and effects. These factors include regulatory elements, such as inspections and market dynamics; organizational features, such as leadership policies and collegial relationships; teachers' professional power and identities; and overlapping policy mandates gaining centrality in global education agendas, such as educational inclusion and innovation. These factors can significantly mitigate or exacerbate accountability effects in the education sector. Gaining a deeper understanding of these factors' role is essential to ensure that assessment and accountability policies enhance education systems rather than undermine teacher autonomy and student learning experiences.

References

References marked with an asterisk indicate documents included in the review data set.

*Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into practice*, *42*(1), 18-29.

*Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System. *Educational Researcher, 37*(2), 65-75. http://www.jstor.org/stable/30137966.

*Amrein, A., & Berliner, D. C. (2002). High-Stakes Testing, Uncertainty, and Student Learning. *Education Policy Analysis Archives,* 10, 1-74. https://doi.org/10.14507/epaa.v10n18.2002

*Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education policy analysis archives*, *18*, 14-14. https://doi.org/10.14507/epaa.v18n14.2010

*Amrein-Beardsley, A., & Holloway, J. (2017). Value-added models for teacher evaluation and accountability: Commonsense assumptions. *Educational Policy*, 33(3), 516-542. doi:10.1177/0895904817719519

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational researcher*, 36(5), 258-267. https://doi.org/10.3102/0013189X0730652

Baert, P. (1991). Unintended consequences: A typology and examples. *International Sociology*, *6*(2), 201-210.

Baker, B. D., Oluwole, J. O., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-Top era. *Education Policy Analysis Archives*, 21(5), 1-71. doi:10.14507/epaa.v21n5.2013

*Basu, R. (2004). Geosurveillance Through the Mapping of Test Results: An Ethical Dilemma or Public Policy Solution?. *ACME: An International Journal for Critical Geographies*, *3*(2), 87-111.

*Béland, D., & Howlett, M. (2016). How solutions chase problems: Instrument constituencies in the policy process. *Governance*, *29*(3), 393-409. https://doi.org/10.1111/gove.12179

*Berkovich, I. (2019). Process implementation perspective on neoliberal regulation: A comparative analysis of national curricula and standards-based reforms in the USA and Australia. *Globalisation, Societies and Education*, 17(5), 593-609. https://doi.org/10.1080/14767724.2018.1559042

Berryhill, J., Linney, J. A., & Fromewick, J. (2009). The Effects of Education Accountability on Teachers: Are Policies Too-Stress Provoking for Their Own Good?. *International Journal of Education Policy and Leadership*, 4(5), 1-14. https://doi.org/10.22230/ijepl.2009v4n5a99

*Bertrand, M., & Marsh, J. (2021). How data-driven reform can drive deficit thinking. *Phi Delta Kappan*, *102*(8), 35-39. https://doi.org/10.1177/00317217211013936

*Bianchi, C., & Salazar, R. (2022). A feedback view of behavioural distortions from perceived public service gaps at 'street-level' policy implementation: The case of unintended outcomes in public schools. *Systems Research and Behavioral Science*, *39*(1), 63-84. https://doi.org/10.1002/sres.2771

*Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas accountability system. *American educational research journal*, *42*(2), 231-268. https://doi.org/10.3102/00028312042002231

*Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2008). The impact of assessment and accountability on teacher recruitment and retention: Are there unintended consequences? *Public Finance Review*, *36*(1), 88-111. https://doi.org/10.1177/1091142106293446

*Buisson-Fenet, H., & Pons, X. (2019). Responsable, mais pas redevable? Gouvernance par les résultats et relations d'"accountability" dans les établissements scolaires en France. *Éducation et sociétés*, *1*, 41-56.

*Bukh, P. N., Christensen, K. S., & Poulsen, M. L. (2022). Performance Funding: Exam Results, Stakes, and Washback in Danish Schools. *Sage Open*, *12*(1). https://doi.org/10.1177/21582440221082100

*Camphuijsen, M. K. (2021). Coping with performance expectations: towards a deeper understanding of variation in school principals' responses to accountability demands. *Educational Assessment, Evaluation and Accountability*, *33*(3), 427-453. https://doi.org/10.1007/s11092-020-09344-6

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, *93*(9-10), 1045-1057. https://doi.org/10.1016/j.jpubeco.2009.06.002

*Cho, J., & Eberhard, B. (2013). When Pandora's Box Is Opened: A Qualitative Study of the Intended and Unintended Impacts Of Wyoming's New Standardized Tests on Local Educators' Everyday Practices. *Qualitative Report*, *18*, 20. https://doi.org/10.46743/2160-3715/2013.1548

Cohen-Vogel, L. (2011). "Staffing to the test" are today's school personnel practices evidence based? *Educational evaluation and policy analysis*, *33*(4), 483-505. https://doi.org/10.3102/0162373711419845

*Collins, C. (2014). Houston, We Have a Problem: Teachers Find No Value in the SAS Education Value-Added Assessment System (EVAAS®). *Education Policy Analysis Archives*, *22*(98), 1-42. https://doi.org/10.14507/epaa.v22.1594

*Cuban, L. (2007). Hugging the middle. Teaching in an era of testing and accountability, 1980-2005. *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas*, *15*, 1-27.

Dahler-Larsen, P. (2014) Constitutive Effects of Performance Indicators: Getting beyond unintended consequences, *Public Management Review*, *16*(7), 969-986. https://doi.org/10.1080/14719037.2013.770058

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers college record*, *106*(6), 1047-1085. http://dx.doi.org/10.1111/j.1467-9620.2004.00372.x

*Datnow, A., & Park, V. (2018). Opening or closing doors for students? Equity and data use in schools. *Journal of Educational Change*, *19*, 131-152. https://link.springer.com/article/10.1007/s10833-018-9323-6

*Davis, T., Bhatt, R., & Schwarz, K. (2015). School segregation in the era of accountability. *Social Currents*, *2*(3), 239-259. https://doi.org/10.1177/2329496515589852

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and management*, *30*(3), 418-446.

*DeMatthews, D. E., & Knight, D. S. (2019). The Texas special education cap: Exploration into the statewide delay and denial of support to students with disabilities. *Education policy analysis archives*, *27*, 2-2. http://dx.doi.org/10.14507/epaa.27.3380

*Demirkasımoğlu, N. (2010). Defining "Teacher Professionalism" from different perspectives. *Procedia-Social and Behavioral Sciences*, *9*, 2047-2051. https://doi.org/10.1016/j.sbspro.2010.12.444

*Diamond, J. B. (2007). Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education*, *80*(4), 285-313. https://doi.org/10.1177/003804070708000401

*Diamond, J. B., & Cooper, K. (2007). The uses of testing data in urban elementary schools: Some lessons from Chicago. *Teachers College Record*, *109*(13), 241-263. https://doi.org/10.1177/016146810710901307

Dolowitz, D.P. and Marsh, D. (2000), Learning from Abroad: The Role of Policy Transfer in Contemporary Policy-Making. *Governance*, *13*, 5-23. https://doi.org/10.1111/0952-1895.00121

*Donaldson, M. L., & Woulfin, S. (2018). From tinkering to going "rogue": How principals use agency when enacting new teacher evaluation systems. *Educational Evaluation and Policy Analysis*, *40*(4), 531-556. https://doi.org/10.3102/0162373718784205

*Ehren, M. C., Gustafsson, J. E., Altrichter, H., Skedsmo, G., Kemethofer, D., & Huber, S. G. (2015). Comparing effects and side effects of different school inspection systems across Europe. *Comparative education*, *51*(3), 375-400. http://dx.doi.org/10.1080/03050068.2015.1045769

*Ehren, M. C., & Hatch, T. (2013). Responses of schools to accountability systems using multiple measures: The case of New York City elementary schools. *Educational*

Assessment, Evaluation and Accountability, *25*, 341-373. http://dx.doi.org/10.1007/s11092-013-9175-9

*Ehren, M. C. & Swanborn, M. (2012) Strategic data use of schools in accountability systems, *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 23*(2), 257-280. http://dx.doi.org/10.1080/09243453.2011.652127

*Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: how public measures recreate social worlds. *American Journal of Sociology, 113*(1), 1–40. http://dx.doi.org/10.1086/517897

Eurydice (2020). *Equity in school education in Europe: Structures, policies and student performance*. Luxembourg: Publications Office of the European Union. https://data.europa.eu/doi/10.2797/658266

Falabella, A. (2020). The Ethics of Competition: Accountability Policy Enactment in Chilean Schools' Everyday Life. *Journal of Education Policy*, *35*(1), 23-45. https://doi.org/10.1080/02680939.2019.1635272

*Feniger, Y., Israeli, M., & Yehuda, S. (2016). The power of numbers: The adoption and consequences of national low-stakes standardised tests in Israel. In *the World yearbook of education 2017* (pp. 15-31). Routledge.

*Fetler, M. (2019). Unexpected testing practices affecting English language learners and students with disabilities under No Child Left Behind. *Practical Assessment, Research, and Evaluation, 13*(1), 6. https://doi.org/10.7275/yz11-h017

*Figlio, D. N., & Getzler, L. S. (2006). Accountability, ability and disability: Gaming the system?. In T. Gronberg and D. Jansen (Eds.), *Improving school accountability* (pp. 35-49). Emerald Group Publishing Limited.

*Fitzgerald, T. (2009). The tyranny of bureaucracy: continuing challenges of leading and managing from the middle. *Educational Management Administration & Leadership*, *37*(1), 51-65. https://doi.org/10.1177/1741143208098164

*Garner, B., Thorne, J. K., & Horn, I. S. (2017). Teachers interpreting data for instructional decisions: Where does equity come in? *Journal of Educational Administration*, *55*(4), 407-426. http://dx.doi.org/10.1108/JEA-09-2016-0106

*Gewirtz, S., Maguire, M., Neumann, E., & Towers, E. (2021). What's wrong with 'deliverology'? Performance measurement, accountability and quality improvement in English secondary education. *Journal of Education Policy, 36*(4), 504-529. http://dx.doi.org/10.1080/02680939.2019.1706103

Greaves, E., Wilson, D., & Nairn, A. (2023). Marketing and School Choice: A Systematic Literature Review. *Review of Educational Research*, *93*(6), 1-37. https://doi.org/10.3102/00346543221141658

Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic reviews, 1*(1), 1-9. https://doi.org/10.1186/2046-4053-1-28

*Haney, W. (2000). The myth of the Texas miracle in education. *Education policy analysis archives*, *8*, 41-41.

*Hargreaves, A. (2020). Large-scale assessments and their effects: The case of mid-stakes tests in Ontario. *Journal of Educational Change*, *21*, 393-420. https://link.springer.com/article/10.1007/s10833-020-09380-5

Harzing, A.W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*, 787–804. https://doi.org/10.1007/s11192-015-1798-9

*Henry, G. T., McNeill, S. M., & Harbatkin, E. (2022). Accountability-driven school reform: are there unintended effects on younger children in untested grades? *Early Childhood Research Quarterly*, *61*, 190-208. https://doi.org/10.1016/j.ecresq.2022.07.005

*Hewitt, K. K. (2015). Educator evaluation policy that incorporates EVAAS value-added measures: Undermined intentions and exacerbated inequities. *Education Policy Analysis Archives*, *23*, 76-76. https://doi.org/10.14507/epaa.v23.1968

*Hinnant-Crawford, B. N. (2023). Legislating instruction in urban schools: Unintended consequences of accountability policy on teacher-reported classroom goal structures. *Urban Education*, *58*(1), 3-35. https://psycnet.apa.org/doi/10.1177/0042085919838004

Hofflinger, A., & von Hippel, P. T. (2020). Missing children: How Chilean schools evaded accountability by having low-performing students miss high-stakes tests. Educational *Assessment Evaluation and Accountability*, 32, 127–152. https://doi.org/10.1007/s11092-020-09318-8

Högberg, B., & Lindgren, J. (2021). Outcome-based accountability regimes in OECD countries: a global policy model? *Comparative Education*, *57*(3), 301-321. https://doi.org/10.1080/03050068.2020.1849614

*Holbein, J. B., & Ladd, H. F. (2017). Accountability pressure: Regression discontinuity estimates of how No Child Left Behind influenced student behavior. *Economics of Education Review*, *58*, 55-67. https://doi.org/10.1016/j.econedurev.2017.03.005

*Holloway, J., & Brass, J. (2018). Making accountable teachers: The terrors and pleasures of performativity. *Journal of education policy*, *33*(3), 361-382. http://dx.doi.org/10.1080/02680939.2017.1372636

Hooge, E., Burns, T. i Wilkoszewski, H. (2012). *Looking Beyond the Numbers: Stakeholders and Multiple School Accountability*. OECD Education Working Papers, *85*. OECD Publishing.

*Horn, I. S. (2018). Accountability as a design for teacher learning: Sensemaking about mathematics and equity in the NCLB era. *Urban education*, *53*(3), 382-408. https://doi.org/10.1177/0042085916646625

Ioannidis, J. P.A. (2023), "October 2023 data-update for "Updated science-wide author databases of standardized citation indicators"", Elsevier Data Repository, *6*, doi: 10.17632/btchxktzyw.6

*Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5-6), 761-796.

Jaffe-Walter, R., & Villavicencio, A. (2023). Leaders' negotiation of teacher evaluation policy in immigrant-serving schools. *Educational Policy*, *37*(2), 359-392. https://doi.org/10.1177/08959048211015614

*Jennings, J. L., & Beveridge, A. A. (2009). How does test exemption affect schools' and students' academic performance? *Educational evaluation and policy analysis*, *31*(2), 153-175.

*Jones, K. L., 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*, *43*(6), 805-822. https://doi.org/10.1080/03054985.2017.1352499

*Kaynak Elcan, N. (2020). A close look at teachers' lives: Caring for the well-being of elementary teachers in the US. *International Journal of Emotional Education, 12*(1), 19-34.

*Kearns, L. L. (2011). High-stakes standardized testing and marginalized youth: An examination of the impact on those who fail. *Canadian Journal of Education/Revue canadienne de l'éducation*, *34*(2), 112-130.

*Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly*, *38*(4), 494-529.

*Landri, P. (2021). To resist, or to align? The enactment of data-based school governance in Italy. *Educational Assessment, Evaluation and Accountability*, *33*(3), 563-580.

*Levatino, A., Parcerisa, L., & Verger, A. (2024). Understanding the stakes: The influence of accountability policy options on teachers' responses. Educational Policy, 38(1), 31-60. https://doi.org/10.1177/08959048221142048

Lingard, B., Martino, W., Rezai-Rashti, G., & Sellar, S. (2015). *Globalizing educational accountabilities*. Routledge.

Lingard, B., Sellar, S., & Lewis, S. (2017). Accountabilities in schools and school systems. In: *Oxford Research Encyclopedia of Education*. Oxford University Press.

Lowenhaupt, R., Spillane, J. P., & Hallett, T. (2016). Education Policy in Leadership Practice: "Accountability Talk" in Schools. *Journal of School Leadership*, 26(5), 783-810. https://doi.org/10.1177/105268461602600503

*Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. *The English Journal*, *91*(1), 79-87.

*Macqueen, S., Knoch, U., Wigglesworth, G., Nordlinger, R., Singer, R., McNamara, T., & Brickle, R. (2018). The impact of national standardized literacy and numeracy testing on children and teaching staff in remote Australian Indigenous communities. *Language Testing*, *36*(2), 265-287. https://doi.org/10.1177/0265532218775758

Maroy, C., & Voisin, A. (2013). As transformações recentes das políticas de accountability na educação: desafios e incidências das ferramentas de ação pública. *Educação & Sociedade*, *34*, 881-901.

Maroy C., & Voisin A. (2017). *Think piece on accountability: Background paper prepared for the 2017/8 Global Education Monitoring Report [Research Report]*. UNESCO. https://halshs.archives-ouvertes.fr/halshs-01705982/document

Maroy, C., & Pons, X. (2019). *Accountability policies in education. A Comparative and Multilevel Analysis in France and Quebec.* Cham: Springer.

Matteucci, M. C., Guglielmi, D., & Lauermann, F. (2017). Teachers' sense of responsibility for educational outcomes and its associations with teachers' instructional approaches and professional wellbeing. *Social Psychology of Education*, 20, 275-298. https://psycnet.apa.org/doi/10.1007/s11218-017-9369-y

*McArthur, D., & Reeves, A. (2022). The unintended consequences of quantifying quality: Does ranking school performance shape the geographical concentration of advantage? *American journal of sociology*, *128*(2), 515-551. https://doi.org/10.1086/722470

Merton, R. K. (1936). The unanticipated consequences of purposive social action. *American sociological review*, *1*(6), 894-904.

*Meadows, M., & Black, B. (2018). Teachers' experience of and attitudes toward activities to maximise qualification results in England. *Oxford Review of Education*, *44*(5), 563-580.

Meadows, D., Richardson, J., & Bruckmann, G. (1982). *Groping in the dark: the first decade of global modelling*. John Wiley & Sons. http://dx.doi.org/10.1080/03054985.2018.1500355

Mintrop, H. (2004). *Schools on probation: How accountability works (and doesn't work)*. Teachers College Press.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery*, 8(5), 336–341. https://doi.org/10.1016/j.ijsu.2010.02.007

*Muñoz-Chereau, B., González, Á., & Meyers, C. V. (2022). How are the 'losers' of the school accountability system constructed in Chile, the USA and England? *Compare: A Journal of Comparative and International Education*, *52*(7), 1125-1144. https://doi.org/10.1080/03057925.2020.1851593

*Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*(2), 263-283.

*Nichols, S. L., & Berliner, D. C. (2005). *The inevitable corruption of indicators and educators through high-stakes testing.* Arizona State University: Tempe

*Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools.* Harvard Education Press.

*Ohemeng, F., & McCall-Thomas, E. (2013). Performance management and "undesirable" organizational behaviour: Standardized testing in Ontario schools. *Canadian Public Administration*, *56*(3), 456-477. https://doi.org/10.1111/capa.12030

*Oyarzún Vargas, G., & Falabella, A. (2022). Indicadores de Desarrollo Personal y Social: La ilusión de la evaluación integral de la calidad. *Psicoperspectivas*, *21*(1), 149-162. https://dx.doi.org/10.5027/psicoperspectivas-Vol21-Issue1-fulltext-2194

Özek, U. (2015); Hold Back To Move Forward? Early Grade Retention And Student Misbehavior. *Education Finance and Policy*, 10(3): 350–377. https://doi.org/10.1162/EDFP_a_00166

Page, M.J., McKenzie, J.E, Bossuyt, P.M, Boutron I, Hoffmann, T.C., Mulrow, C.D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372(71). https://doi.org/10.1136/bmj.n71

*Pagès, M. (2021). Enacting performance-based accountability in a Southern European school system: between administrative and market logics. *Educational Assessment, Evaluation and*

Accountability, 33, 535–561. https://link.springer.com/article/10.1007/s11092-021-09359-7

*Park, V., & Datnow, A. (2017). Ability grouping and differentiated instruction in an era of data-driven decision making. *American Journal of Education*, *123*(2), 281–306. https://doi.org/10.1086/689930

Pawson, R., Greenhalgh, T., Harvey, G., & Walshe, K. (2005). Realist review—A new method of systematic review designed for complex policy interventions. *Journal of Health Services Research & Policy*, 10, 21–34. https://doi.org/10.1258/1355819054308530

*Penninckx, M., Vanhoof, J., De Maeyer, S., & Van Petegem, P. (2016). Enquiry into the side effects of school inspection in a 'low-stakes' inspection context. *Research Papers in Education*, *31*(4), 462-482. https://doi.org/10.1080/02671522.2015.1076886

*Perryman, J., Ball, S., Maguire, M., & Braun, A. (2011). Life in the pressure cooker–School league tables and English and mathematics teachers' responses to accountability in a results-driven era. *British Journal of Educational Studies*, *59*(2), 179-195.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford: Blackwell Publishing. https://doi.org/10.1002/9780470754887

*Polesel, J., Rice, S., & Dulfer, N. (2014). The impact of high-stakes testing on curriculum and pedagogy: A teacher perspective from Australia. *Journal of education policy*, *29*(5), 640-657. https://doi.org/10.1080/02680939.2013.865082

Polikoff, M. S. (2012). The association of state policy attributes with teachers' instructional alignment. *Educational Evaluation and Policy Analysis*, 34(3), 278-294.

Popper, K. (1944). The poverty of historicism, II. A Criticism of Historicist Methods. *Economica*, *11*(43), 119-137.

Portz, J., & Beauchamp, N. (2022). Educational accountability and state ESSA plans. *Educational Policy*, *36*(3), 717-747. https://doi.org/10.1177/0895904820917364

*Rezai-Rashti, G. M., & Segeren, A. (2020). The game of accountability: perspectives of urban school leaders on standardized testing in Ontario and British Columbia, Canada. *International Journal of Leadership in Education*, *26*(2), 1-18. http://dx.doi.org/10.1080/13603124.2020.1808711

*Russell, J. L., & Bray, L. E. (2013). Crafting coherence from complex policy messages: Educators' perceptions of special education and standards-based accountability policies. *Education Policy Analysis Archives*, *21*, 1-22. https://doi.org/10.14507/epaa.v21n12.2013

*Rustique-Forrester, E. (2005). Accountability and the pressures to exclude: A cautionary tale from England. *Education Policy Analysis Archives*, *13*(26), 1-41.

*Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary educational psychology*, 61, 101860.

Sartre, J. P. (1960). Critique de la raison dialectique/1 Théorie des ensembles pratiques. *Critique de la raison dialectique*.

*Scott, C. M. (2017). Un-"chartered" waters: Balancing Montessori curriculum and accountability measures in a charter school. *Journal of School Choice*, *11*(1), 168-190. http://dx.doi.org/10.1080/15582159.2016.1251280

*Shirrell, M. (2018). The effects of subgroup-specific accountability on teacher turnover and attrition. *Education Finance and Policy*, *13*(3), 333-368. https://doi.org/10.1162/edfp_a_00227

*Skedsmo, G., Rönnberg, L., & Ydesen, C. (2020). National testing and accountability in the Scandinavian welfare states: Education policy translations in Norway, Denmark and Sweden. In: *World yearbook of education* 2021 (pp. 113-129). Routledge.

Spreen, C-A. (2004). Appropriating borrowed policies: Outcomes-based education in South Africa. In G.Steiner-Khamsi (Ed.), *The global politics of educational borrowing and lending* (pp. 101–113). Teachers College Press.

*Thiel, C. (2021). Side effects and the enactment of accountability: results of a comparative study in two German federal states. *Educational Assessment, Evaluation and Accountability*, *33*, 403-425. https://link.springer.com/article/10.1007/s11092-021-09358-8

*Thiel, C., & Bellmann, J. (2017). Rethinking Side Effects of Accountability in Education: Insights from a Multiple Methods Study in Four German School Systems. *Education Policy Analysis Archives*, *25*(93), 1-32. http://dx.doi.org/10.14507/epaa.25.2662

Tse, T. K. C. (2019). Fears and Tears of Transparency and Disclosure: Controversies and Politics of School Profiles in Hong Kong Since 2000. *Education and Urban Society*, *51*(8), 1106-1126. https://doi.org/10.1177/0013124518785014

*Van Wyk, M., & Le Grange, L. (2016). Die geleefde ervarings van primêre skoolonderwysers binne'n kultuur van performatiwiteit: navorsings-en oorsigartikels (2). *Tydskrif vir Geesteswetenskappe*, *56*(4-2), 1149-1164. http://dx.doi.org/10.17159/2224-7912/2016/v56n4-2a4

*Vasquez Heilig, J. (2011). Understanding the interaction between high-stakes graduation tests and English learners. *Teachers College Record*, *113*(12), 2633-2669. http://dx.doi.org/10.1177/016146811111301209

*Vasquez Heilig, J., Young, M., & Williams, A. (2012). At-risk student averse: risk management and accountability. *Journal of Educational Administration*, *50*(5), 562-585. https://doi.org/10.1108/09578231211249826

Verger, A., Fontdevila, C., & Parcerisa, L. (2019a). Reforming governance through policy instruments: How and to what extent standards, tests and accountability in education spread

worldwide. Discourse: Studies in the Cultural Politics of Education, 40(2), 248-270. https://doi.org/10.1080/01596306.2019.1569882

Verger, A., Parcerisa, L., & Fontdevila, C. (2019b). The growth and spread of large-scale assessments and test-based accountabilities: A political sociology of global education reforms. Educational Review, 71(1), 5-30. https://doi.org/10.1080/00131911.2019.1522045

Vernon, R. (1979). Unintended consequences. *Political theory*, *7*(1), 57-73.

Watanabe, M. (2008). Tracking in the era of high-stakes state accountability reform: Case studies of classroom instruction in North Carolina. *Teachers College Record, 110*(3), 489-534. https://doi.org/10.1177/01614681081100030

*Whitney, C. R., & Candelaria, C. A. (2017). The effects of No Child Left Behind on children's socioemotional outcomes. *AERA Open*, *3*(3), 1-21.

*Wiggins, A., & Tymms, P. (2002). Dysfunctional effects of league tables: a comparison between English and Scottish primary schools. *Public money and management*, *22*(1), 43-48. https://doi.org/10.1111/1467-9302.00295

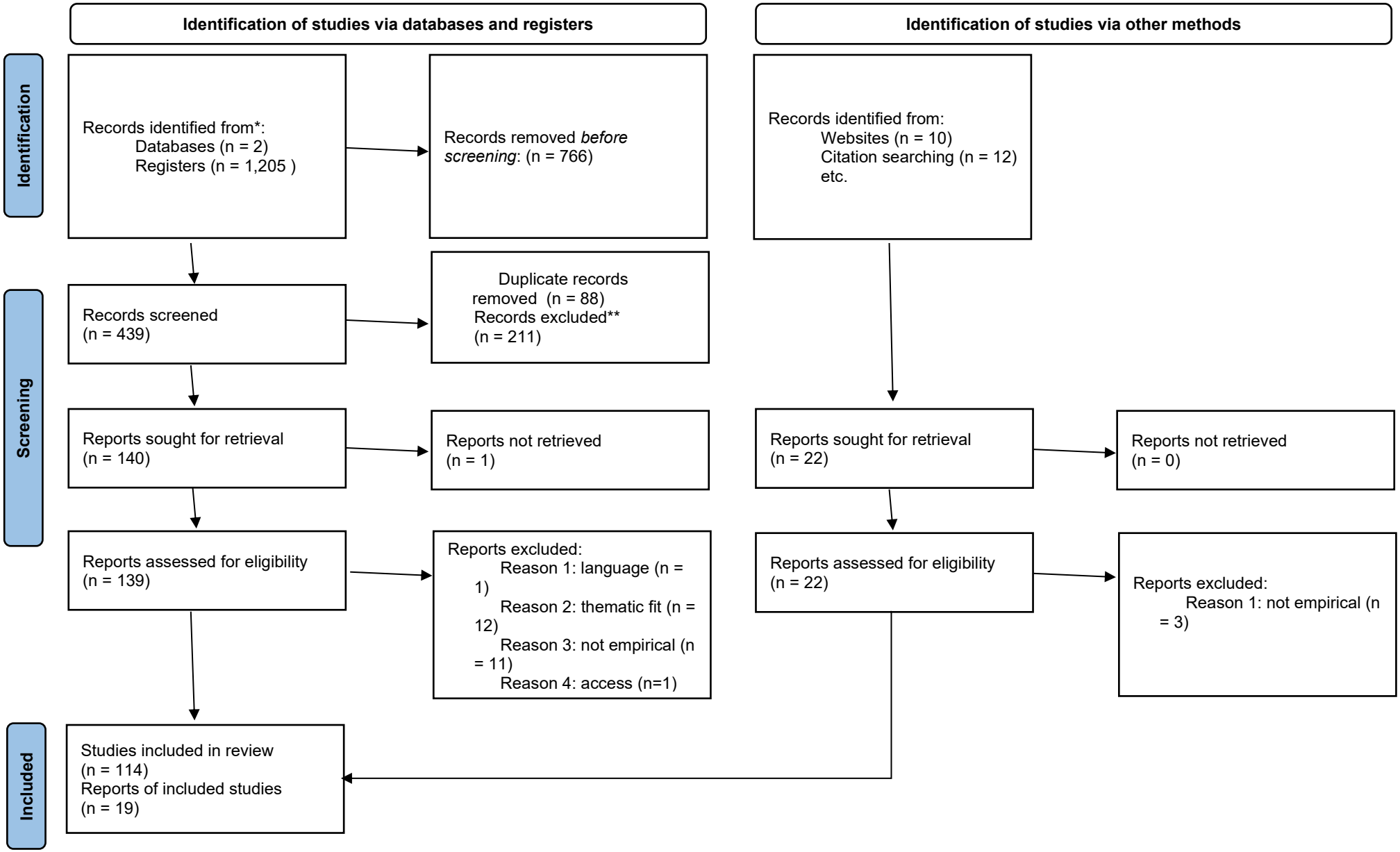
*Wills, J. S., & Sandholtz, J. H. (2009). Constrained professionalism: Dilemmas of teaching in the face of test-based accountability. *Teachers college record*, *111*(4), 1065-1114. https://doi.org/10.1177/016146810911100401

*Wilson, D., Wilson, D., Croxson, B., & Atkinson, A. (2006). "What gets measured gets done" Headteachers' responses to the English secondary school performance management system. *Policy Studies*, *27*(2), 153-171., B., & Atkinson, A. (2006). "What gets measured gets done" Headteachers' responses to the English secondary school performance management system. *Policy Studies, 27*(2), 153-171.

Zhao, Y. (2017). What works may hurt: Side effects in education. *Journal of Educational Change*, 18(1), 1-19. https://doi.org/10.1007/s10833-016-9294-4

Zhu, J., & Liu, W. (2020). A tale of two databases: The use of Web of Science and Scopus in academic papers. *Scientometrics*, *123*(1), 321–335.

Figure 1.
PRISMA 2020 flow diagram with searches of databases, registers and other sources



Source: Adapted from Page et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372(71). doi: 10.1136/bmj.n71. For more information, visit: http://www.prisma-statement.org/

Appendix

Table 1. Search terms used in database searches

Education search terms	**Operator**	**PBA search terms**	**Operator**	**Side-effects search terms**	**Operator**	**Years**	**Operator**	**Excluded subject areas**
School*, teacher, school principal*, head teacher*	AND	Accountability, new public management, school-based management, benchmarks, teacher* evaluation, testing, league table, ranking, high stake*, low stake*	AND	Undesired, unexpected, unintended, side effect*	LIMIT TO	2000-2022	Exclude	Mathematics, medicine, computer science, agricultural and biological sciences, environmental science, health professions

Data analysis

The codebook presented below was elaborated in an iterative manner and includes the following two dimensions:

- *Descriptive data about the investigation*: These data identify the conceptual framework, the main research questions, aims, and methods of the study. It also includes codes to assess the overall quality of the research.

- *Substantive issues about the emergence of side-effects*: This dimension includes codes to analyze the context of the policy intervention (school level, school characteristics, country), the policy design, the mechanisms, and the outcomes (namely, external contexts, professional, pedagogical, organizational and students side-effects).

During the analyiss, the articles were distributed among a team of six researchers that performed a qualitative content analysis of the documents included in the final dataset. The distribution of articles was organized on the basis of accountability regimes (we distinguish between thick and thin accountability policy models on the basis of Euryidice (2020) and Högberg & Lindgren (2021).

Table 2. Codebook

Code	Definition	Code group
RQ and objectives	Identify the main research questions and the aims of the study	Conceptual framework, research design and methodology
Conceptual framework	Identify the main theories used in the paper. Please, combine this code with emerging codes that allow identifying the theory (e.g., Policy enactment theory, sense-making theory, etc.)	
Key references		
Quantitative methods		
Qualitative methods		
Mixed methods		
Definition of side-effects	Key definition(s) of what is a side-effect, according to the paper.	
School level	Primary, secondary.	Context
School characteristics	History of the school, students' background and schools' composition (e.g., elite school, middle class, disadvantaged schools), leadership characteristics	
Region/country	Please, combine this code with in-vivo code with the name of the country (e.g., England). This code will be used as proxy of educational system characteristics (e.g., marketisation levels, and so on).	
Mechanism	Highlighting the specific, explicit causation described in the paper, between policies, contexts, and outcomes. Acknowledging that not all papers would have this explicit relationship and that in some cases we will have to identify this causal explanation ex post.	Mechanims
Students' side-effects	Impacts on students' conditions (e.g., stress and anxiety, well-being, etc.).	Outcomes/side-effects

Pedagogical side-effects	E.g.. teaching to the test; aligning curriculum and instructions' strategies to the tests; cheating; triage; etc.	
Organizational side-effects	E.g. reshaping the test pool; reallocation of time and resources; putting 'best' teachers to teach the test; etc.	
Professional side-effects	E.g. changes in teachers' identity; changes in teachers' preparation; changes in professional environment (e.g., stress, burnout); reputational concerns, pressure to perform; etc.	
External contexts side-effects	The impact of accountability policies and instruments (e.g., test' results, league tables, schools' rankings, etc.) on local education markets' dynamics (i.e., the position of the school in relative comparison to other schools) and families' perceptions and behavior (i.e., how those results, tables and rankings are perceived by families and used in school choice processes). Also: on how neighborhoods and minorities are perceived/increasing segregation-inequality.	
Theory of change	Description of the theory of change of the accountability policy (or the purpose of the test, e.g. equity purposes, transparency, control, increasing quality of education, efficiency).	
Material consequences	E.g., salary bonus, school closure, students' graduation; teachers´ promotion decision; school rewards..	
Symbolic consequences	Merged from Reputational consequences and Symbolic consequences	Policy design
Professional/pedagogical consequences	E.g., development through professional courses, pedagogical support.	
Collective consequences	The policy consequences have affected all teachers or the school	
Individual consequences	The policy design includes individual consequences for the principal or individual teachers	

High-stakes accountability regimes	(although it may be derived from financial, material incentives it is still a useful code for easily grouping papers).	
Low-stakes accountability regimes	Accountability systems with soft consequences (reputational, pedagogical, etc.)	
Account-holder	Who is the principal? i.e., who is imposing the consequences (e.g., Ministry of education, etc.).	
Account-giver	Who is the actor? i.e., the actor receiving the consequences (who is evaluated).	
Test characteristics	E.g. competence-based, curriculum-based; level; used for advancing course. (if possible); characteristics of the performance measures and how test results are reported	
Quality-middle	The document refers to relevant literature and reduces the ideological bias by referring to different approaches. The document has a solid conceptual framework (central concepts are defined or their original sources are conveniently referred), but it oversimplified some of the concepts or makes inaccurate inferences from primary or secondary data. The methodology used in the research is made explicit. (Verger et al. 2016, p. 199)	Quality
Quality-high	The document is based on relevant literature and a rigorous conceptual framework. It reaches logical and reasoned conclusions, providing broad evidence. The theoretical framework and the final conclusions are clearly and properly linked. The methodology used is made explicit and adequately employed (Verger et al. 2016, p. 199)	

Quality-low	The document does not mention relevant literature to support its statements, or there are not references to earlier work in or significant contributions to the field. The document does not clearly define the concepts. The document draws conclusions without providing relevant evidence; inferences are unsupported. (Verger et al. 2016, p. 199)