Research Article

EDUCATIONAL POINT
An International Peer-Reviewed Journal

https://www.educationalpoint.net
ISSN: 3062-1011 (Online)

Schools' Evaluation Drift: Inconsistencies and Interpellations of a High-Stakes Inspection System

Lídia Serra [1*], José Alves [1]

[1] Universidade Católica Portuguesa, PORTUGAL
* Correspondence: lidiajpserra@gmail.com & ljserra@ucp.pt

CITATION: Serra, L., & Alves, J. (2025). Schools' Evaluation Drift: Inconsistencies and Interpellations of a High-Stakes Inspection System. *Educational Point*, *2*(2), e139. https://doi.org/10.71176/edup/17660

ARTICLE INFO

Received: 27 September 2025
Accepted: 20 December 2025

OPEN ACCESS

ABSTRACT

A consensus exists in transnational educational policy regarding the relevance of accountability and the contribution of school evaluation to education quality. This paper scrutinises the evaluation results of Portuguese schools provided by the Inspectorate services using a pairwise comparison between 194 schools evaluated in 2018-2020 and 2021-2023. Regarding the literature gap on the behaviour of accountability systems over time, this study can contribute to a reflection on justice and transparency in high-stakes systems. The findings suggest that (i) the external evaluation results drift following a directional evolutionary model, indicating progression concerning the self-evaluation, educational services, and results domains; (ii) the occurrence of standardisation and leadership legitimisation phenomenon; (iii) possible side effects in the schools' evaluation process, namely evasive behaviour, apparent and constructed realities, and evaluation distortion; (iv) the external evaluation framework flexibility in accommodating territorial differences between schools without producing system disadvantage. Departing from insights into how a high-stakes external evaluation system operates over time, the study offers an empirically grounded assumption that reveals dynamics not unique to Portugal, but characteristic of accountability regimes adopted across many educational systems. In conclusion, to improve the quality of education, low-stakes accountability systems should be implemented to strengthen transparent schools' autonomy.

Keywords: accountability, external evaluation, high-stakes system, legitimation, trust, organised hypocrisy

INTRODUCTION

Accountability is a transnational travelling concept considered a powerful policy tool in education. Quality and school improvement narratives are yielded at the expense of control and regulation mechanisms, and standardisation appears to facilitate auditing and evaluation (Mufic, 2023). The growing accountability in the educational systems over the last 30 years has brought pressure and trust problems into the school practice (Hanberger et al., 2016). High-stakes accountability systems are allocated to a market-oriented philosophy (West & Pennell, 2005), and are implemented through student tests operated as high-stakes instruments (Cavendish et al., 2017; Heilig et al., 2018), high-stakes teacher accountability policies due to teachers' central role in the educational change (Pizmony-Levy & Woolsey, 2017), and as high-stakes inspections in scenarios of accountability-based governance (Liu et al., 2021). High-stakes accountability systems are tools designed to feed up evaluative diagnosis to make decisions concerning the school organisation, the school actors, and the pedagogical priorities (Constantinides, 2022; Ólafsdóttir et al., 2022). Accountability systems act as pressure to ensure and control quality improvements, and as a significant driver for changes (Liu et al., 2021) with implications for teachers' careers (Cavendish et al., 2017; Pizmony-Levy & Woolsey, 2017).

Evaluating the quality of education, enhancing student outcomes, and refining teachers' instructional practices are key directives of educational policy. External accountability that supports strong dialogical internal accountability and emphasises capacity-building for improvement (Cochran-Smith, 2021) should be driven collaboratively with teachers and school leaders in a self-improving system, which may lead to quality in both processes and outcomes (Hutt & Lewis, 2021). In the opposite direction, an "accountability system that follows economic features and induces mistrust, perverts transparency, and blocks school improvement" (Cochran-Smith, 2021, p.15) can compromise the purpose of quality. Campbell (1979) discussed the problem of effects or distortions due to pressure exerted on systems. Hence, we depart from the premise that high-stakes inspection systems are vulnerable to distortion through pressure to hold school leaders and teachers accountable for student performance. Given that how an evaluation system shapes and responds during an ongoing evaluative cycle is not well understood, a question arises: *Is it stable, or does it adjust? Does it ensure trust and justice in the system and outcomes?*

In Portugal, several research studies concerning schools' external evaluation, departing from the national educational inspectorate services action, have been developed, including on pedagogical supervision practices (Seabra et al., 2021), the perspective of teachers collaboration (Seabra et al., 2022; Miranda et al., 2023), improving areas and strengths regarding the schools' educational strategy and action (Serra et al., 2023a), educational changing mechanisms and innovation (Barreira et al., 2023; Serra et al,, 2025), education quality (Oliveira, 2017), effects on the schools autonomy (Mouraz et. al., 2019), and regulation mechanisms (Carvalho & Costa, 2017). There is diversity in research on external evaluation in the Portuguese context, similar to the international landscape, in which accountability systems, according to Fahey and Köster (2019), have been extensively scrutinised and are essential components of strategic governance in complex education systems. In contrast, longitudinal studies on external evaluation are scarce and can contribute to a better understanding of the level of stakes in accountability systems.

In light of the exposed literature gap, this article presents an analytical exploratory exercise using results from the Portuguese schools' auditing conducted over an evaluative cycle to infer the existence of pressure, distortion, and bias in the responsibility system. After presenting a theoretical perspective about accountability purposes in educational systems, the article seeks responses regarding the ongoing Portuguese profile of the schools' external evaluation to discuss if (i) it affords trust, consistency, and transparency; (ii) it is susceptible to political conditioning; and (iii) it allows transparent and sustainable processes of educational quality improvement. In this sense, the paper presents an empirical study that uses inferential statistical analyses to problematize the consequences, threats, and risks associated with the external evaluation profile of Portuguese schools. The paper concludes by raising considerations for upgrading educational accountability

systems that may improve quality and avoid the rituals of an 'organised hypocrisy' that Brunsson (2006) pointed out.

THEORETICAL FRAMEWORK

Educational Accountability and the Political Agendas

Testing, examining, assessing, auditing, evaluating, surveying, and monitoring are long-standing prerogatives of the educational system at the expense of quality and intelligence-based accountability assumptions. At the beginning of the 21st century, contradicting the logic of an evaluation focused on the schools' compliance with norms, accountability shifted to focus on students' performance. However, nowadays, it is definitely changing to an all-system approach, and according to Portz (2021), to a multiple metrics process.

"Lack of accountability and control in schools is the primary underlying issue related to low-performing teachers" (Küçükbere & Balkar, 2021, p.168). Professional and institutional responsibility requires schools, leaders, and teachers to become accountable to society and governments. However, the illusion that high-stakes accountability policies are needed and possible (Pizmony-Levy & Woolsey, 2017) and the perspective of a staged control (Afonso, 2015) to ensure social confidence requires a new generation of accountability. Intelligent accountability (Lillejord, 2020; O'Neill, 2013) and a more robust system (Portz, 2021) shaped by cultures of continuous improvement engaged in generating professional capital, organisational capital, and innovation, are critical elements for improving true organic interdependency in schools (Serra et al., 2023b). Accountability is a powerful mechanism for regulation and improvement, serving quality, political, socio-economic, legitimisation, and standardisation purposes.

Quality Purposes

"Accountability is considered a mechanism that ensures the professional development of teachers and thus improves professional performance" (Küçükbere & Balkar, 2021, p. 168), as well as leadership and the strategic planning defined for the school. The schools' external and self-evaluation are the nexus of the principals' leadership and professionalism, and accountability is a central element for building the schools' organisational and the teachers' professional capital. New forms of organisation that ensure the school's responsiveness, interdependence, and value of decisional, social, and human capital are necessary for improvement (Fullan et al., 2015; Hargreaves & Fullan, 2012; Serra et al., 2023a). All school capital can be nurtured by an organised, consistent, deep, and responsive mechanism of school knowledge, namely external evaluation, but especially self-evaluation (Serra et al., 2024). In practice, the school evaluation follows a diptych mechanism: (i) one, a constructive, combined, and complementary action between external and internal schools' evaluation (Donaldson, 2013) significantly supportive; (ii) other, more dispersive, disconnected, and dissociative regarding both instruments of the schools' evaluation, especially in high-stakes systems which produce side-effects due to accountability pressures (Penninckx, 2017). Internal accountability has been identified as the primary strategic supportive process of school guidance, and according to Fullan et al. (2015), it should precede external accountability. Then, self-evaluation appeared as an instrument for external evaluation purposes (OECD, 2020), and inspectorates now emphasise it (Simeonova et al., 2020).

External school evaluation has too many deficits and needs to be intertwined with school-based self-evaluation (McNamara & O'Hara, 2008; Penninckx, 2017). School evaluation provides conceptual and contextual knowledge for principals, middle leaders, and teachers to assume professionalism, meaning responsibility regarding student learning and professional development. Accountability and school evaluation provide organisational awareness and, if virtuously conceived, can have a feed-forward effect that shapes subsequent policies.

Political and Social-Economic Purposes

An accountability system should occur along the "axis of professionalism, or a harmful rather than supportive system will emerge regarding occupational professionalism" (Küçükbere & Balkar, 2021, p. 168). Promoting professional interactive development based on trust is crucial (Hargreaves, 2019), as the absence of it can lead to concealment dynamics, individualism, conservatism, and balkanisation. Portz (2021) debated the issue of trust in relation to the consequences of the external evaluation process, specifically regarding rewards and punishments based on metric reviews relative to established goals. Pushing the educational system to accountability and imposing high levels of scrutiny can degenerate the purposes of pursuing quality education. Hutt and Lewis (2021) refer to the aspiration to dodge accountability structures built around high-stakes performative measures. It can depreciate trust (Zamir, 2019) and transparency (Roberts, 2018), and collide with a self-improving system. Cochran-Smith (2021) suggests that:

> "Accountability, on its own, is neither good nor bad, but rather depends on the larger policy and political agendas to which it is attached, how it is used, the goals, values, and purposes it serves, and the assumptions it makes about who should be accountable for what, to whom, and for what purposes" (p. 8).

The "risk of the quality discourse will equate 'quality' with the fulfilment of quality criteria", meaning that the "quality work will no longer be a goal to strive for", drifting to "checkpoints that should simply be achieved and warning flags should be lowered, driven by a fear of being punished for failures" (Mufic, 2023, p. 647). Addressing punishments and rewards in external evaluations for saving purposes is common in countries with fragile economies, such as Portugal's. In Portugal, the external evaluation determines the quotas available for teachers' and principals' career progression, which can lead to a degeneration of the primary purpose of improving the system quality. The risks of hindering higher professional commitment and teachers' deep and wide contributions are possible previews. This context of pressure is further exacerbated by national exams that influence schools' national rankings. "These external forms of accountability have become increasingly 'high stakes' given that a school's reputation is based on its performance on these measures" (Keddie, 2015, p. 2). High-stakes-testing programmes (and high-stakes inspection mechanisms) add pressure on school leaders, which may pressure teachers instead of providing more supportive supervision, and they can collide with the quality and nature of teachers' contributions (Six, 2021).

Campbell's law asserts that "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1979, p. 85). Side effects are frequently observed in accountability systems (Thiel & Bellmann, 2017), and fabricating schools' images and those of their leaders (Afonso, 2015) can undermine the effectiveness of inspections. Although the purpose of quality is rhetorically unquestionable in the school's external evaluation, the literature describes some controversial effects. Some studies reunited information regarding external evaluation on the positive impact on students' performances and school change (Ehren & Visscher, 2006; McCrone et al., 2009). Constructive external evaluation reports are vital for teachers and principals to identify areas for improvement and to implement many of the recommended changes (Ólafsdóttir et al., 2022).

Other studies suggest that external evaluations had little value in helping teachers improve their practice, inducing anxiety (Hanberger et al., 2016), frustration due to fixation on results and problems (Schillemans & Bovens, 2011), cultures of silence and stress (Hanberger et al., 2016), focus on quantitative results (Dahler-larsen, 2014), reduce trust, inhibit discussion, and diminish honest self-evaluation (Hopkins et al., 2016), processes of staging during auditing (Afonso, 2015), and eroding commitment to professional responsibility (Matteucci et al., 2017). Other studies suggest that the induced changes provided by external evaluation depend more on the school's cultures, reuniting evidence that after evaluation, schools are changing at

different velocities and with some schools still struggling to find a way of integrating external evaluation orientations or exhibiting a lack of understanding and adherence to the orientations (Serra et al., 2024). These controversial results support that trust-based accountability is sustained in compliance with "trust being part of relatedness directly impacts the support for psychological needs and internalisation of values" (Six, 2021, p. 71).

Thiel and Bellmann (2017) found that side effects are inherent to high-stakes contexts but are also a substantial problem in low- or no-stakes contexts. The same authors suggest that while most teachers and principals evidence adaptive behaviour in the no- and low-stakes contexts, accountability regimes with a high degree of market and bureaucratic pressure seem to foster evasive behaviour, meaning marks of staging or organised hypocrisy. Hence, the challenge is to develop an accountability system that operates through the axis of professionalism; otherwise, a harmful rather than supportive system will emerge regarding occupational professionalism (Küçükbere & Balkar, 2021).

Legitimisation Purposes

The primary function of the evaluation systems is "to support and legitimise local governance by objectives and results, through monitoring and evaluating schools and student performance and using performance results to take action to improve student performance and equity" (Hanberger et al., 2016, p. 360). Evaluation systems evidence a symbolic effect, which, according to Penninckx (2017), is translated into the extent to which the inspection report is treated as an opportunity to legitimise an opinion already held by the inspection. Another point is to increase the legitimacy of the school assignment in the eyes of society by providing some assurance of procedural fairness (Sattin-Bajaj & Jennings, 2020), and simultaneously making the educational system accredited.

The school organisation allocates "challenges faced by principals in exerting legitimate power", which give rise to "strategies deployed by the principals to use legitimate power effectively" (Sorm & Gunbayi, 2018, p. 267). Studies reveal that leadership authority is not innate; instead, it is constructed over time. Then, leadership must be continuously sustained through social interaction, with the discursive process of legitimation playing a central role in securing others' acceptance of the leader's power (Chin et al., 2019). Legitimation of the leader-follower relationship is expressed in external evaluations as a more favourable appreciation of leadership and management compared to other domains, and sometimes with problematic levels of consistency in leadership activity (Serra et al., 2023a). According to Barroso (2022), in times of prevailing models of school autonomy, this form of regulation represents a legitimising rationality developed at the expense of a posteriori control through external evaluation and processes aligned with new public management.

Standardisation Purposes

More outstanding local autonomy increases actors' power to self-organise, collaborate, support innovation, and reduce the tendency toward compliance (Cochran-Smith, 2021). However, a higher level of scrutiny can impair the quality of accountability and standardise schooling through a top-down process of hierarchical control (Hanberger et al., 2016). The reduction of autonomy through standardising the quality of work takes shape over time, as new teachers are educated in performative systems and come to take external accountability standards for the quality of their work for granted (Ehren & Bachmann, 2020). Constitutive and "tacit or indirect effects, for example, how evaluation (systems) can shape discourses, defining what is important in education and school systems" (Hanberger et al., 2016) can also standardise good schooling. Narrow standardised emergent responses can reduce professional agency and limit the "heterogeneous nature of education, where teachers (and schools) need a degree of professional discretion to meet individual learning needs", making "such monitoring highly contentious" (Ehren & Bachmann, 2020, p.40-41). Six (2021) suggests that such predictability typically comes from the inspection frameworks, which standardise how schools are inspected and the criteria by which their quality is assessed. However, if intelligent and balanced

accountability is developed and autonomy is exercised, "these combinations can repair some of the flaws inherent in each of the ideal types, such as a lack of innovation and variety in over-standardised hierarchies, high inequality in (quasi)markets or groupthink, ambiguity and high transaction costs in networks" (Ehren & Bachmann, 2020, p. 51).

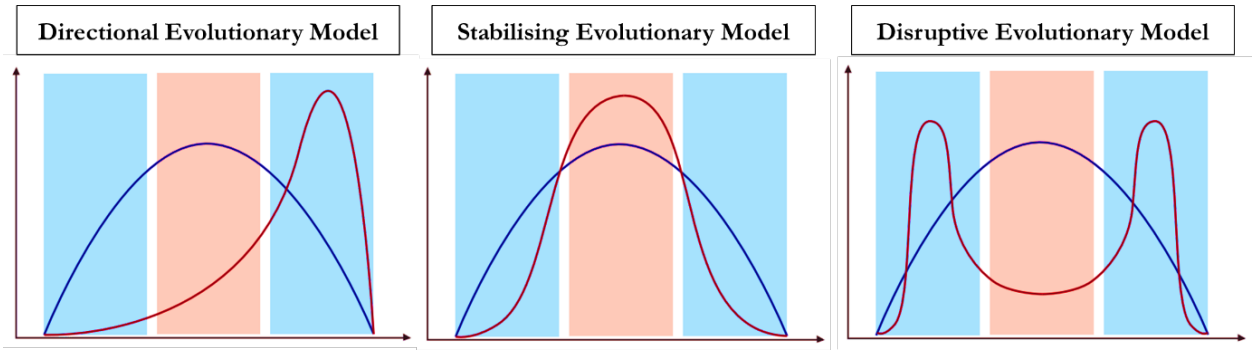
Despite the growing autonomy desideratum for education, paradoxically, the deregulation and re-regulation dichotomy can lead to forms of state control where "schools and professionals are becoming more accountable for providing education and achieving results" (Helgøy et al., 2007, p.199). Indeed, the external evaluation effects depend on the extent to which inspection "influences the thinking of decision-makers and as such may have an impact on their actions" (Visscher & Coe, 2003, p. 58). "Accountability should be reimagined and reclaimed through the powerful alternative paradigm of a democratic accountability in teacher education" (Cochran-Smith, 2021, p. 12), because "school evaluation systems that are poorly designed can turn the evaluation process into a punitive, compliance-based exercise that is unrelated to schools' core activities" (OECD, 2020, p. 4).

Accountability and Trust: A Context for the Problem

As a bond of the educational system, the schools' evaluation process provides the necessary external glance to align and reorient the schools' building identity process. It can enrich the basis for promoting and assuring a collective sense of vision and mission to develop a school project of inclusion and equity in a context. "Intelligent accountability is grounded in trust and capacity building, and it is deliberately organised to yield information" to sustain a "thoughtful programme improvement, rather than used primarily to demonstrate that a particular regulator or other agency is doing its job by holding teacher education accountable" (Cochran-Smith, 2021, p.14). The inspection frameworks define criteria for the schools' evaluation, and "these standards allow schools to understand how they are being evaluated and when, as well as how they can prepare for an inspection, preventing unpleasant surprises and a feeling of unfairness for being judged on criteria that were unknown to then" (Six, 2021, p. 117). However, especially in high-stakes accountability systems, it may lead to the *cat-and-mouse* effect, defined in the English language dictionary as a *contrived action involving constant pursuit, near captures, and repeated escapes.* In staging processes, the parameterised external evaluation framework may produce mimicry effects in the school organisations.

Trust and pressure, underlying features of an accountability system, can shape responses in the school's organisation regarding inspection acts. These responses can model organisational and pedagogical improvements regarding Inspectorate orientations and pronounce the logic of the *cat-and-mouse* effect. Portugal follows a high-stakes inspection system and utilises external evaluation for quality purposes linked to bonuses and penalties. In such a case, the question of 'how the system reacts over time and if the schools' evaluation is changing?' is mandatory. **Figure 1** presents the study model, which elucidates three hypothetical behaviours of the school's evaluation system over time.

The stabilising model occurs when the normal distribution of the schools' evaluations results remains constant over time, and an alignment with a preview mid-fix point is observed. In this model, the distribution of ongoing external evaluation results ripens the 'natural' normal distribution and expresses its intrinsic variability. The directional model shows an over-time drift towards one end of the distribution, indicating that the evaluation is not stabilised but is undergoing transformation. Time and environmental pressures (due to social, political, and inspection inputs), as well as contextual factors (resulting from internal school organisation in response to local mutations and environmental pressures), influence the system and drive its evolution. The disruptive model applies to situations in which the population splits into two due to specific territorial variations in the social tissue and eventually becomes a target of macro-level political decisions and programs. Examples of this kind of intervention are programs worldwide regarding social inclusion that use education as a tool, like



Note: blue line – preview distribution of the population; red line – population distribution after a given period.

Figure 1. Model of study

the 'education action zones' in England, and similarly, the 'zones d'éducacion prioritaire' in France, and 'territórios educativos de intervenção prioritária' (TEIP) in Portugal.

In Portugal, the schools' external evaluation initiated a third cycle in 2018 with a new framework that introduced a new domain – self-evaluation – in addition to the previous domains: leadership and management, educational services, and results. This new cycle included lesson observations, added to the last schedule, which encompassed a survey of the community (students, parents, and staff), school documents analysis, visits to facilities, a presentation session of the school organisation, panels with specific organisational groups, and data analyses of outcomes provided by the statistics information system of the Education Ministry. Considering the background information discussed and the assumption that a trusty external evaluation following quality purposes is catching the whole system heterogeneity, we hypothesized that:

H1**:** The Portuguese external evaluation of schools follows a stabilising evolutionary model.

In Portugal, political responses to inclusion and equity led to the creation of the TEIP program in 1996, prioritising interventions in schools in disadvantaged areas with poor academic performance. Even though the implementation of positive discrimination policies, the contextual differences regarding schools outside these vulnerable zones lead us to the hypothesis:

H2: *Following a disruptive evolutionary model, the Portuguese external evaluation of TEIP and non-TEIP schools differs*.

METHOD

Data Collection

This study utilised the external evaluation reports produced during the third cycle of the school's evaluation by the Portuguese Inspectorate Services as a documentary corpus. All reports are available at an online public repository. This cycle began in 2018, and we employed a saturated sampling procedure, considering all external evaluation activities provided up to 2023, a total of 194 external evaluation reports. In each cycle, each school or school cluster (hereafter referred to as a school) was evaluated only once. The study analysed every public school's external evaluation reports, excluding those from the pilot phase of the external evaluation cycle, as well as professional, artistic, and private schools. We analysed 194 schools' evaluation reports according to four domains defined in the Portuguese Inspectorate Services evaluation framework: schools' self-evaluation, leadership and management, educational service, and results. The reports included a qualitative evaluation of a school's standard performance in each domain, rated on a five-point scale assigned officially by the Inspectorate.

We coded each domain's qualitative evaluation on a five-point scale (1 = insufficient, 2 = sufficient, 3 = good, 4 = very good, and 5 = excellent). For sample characterisation purposes, we reunited information concerning the number of schools within each cluster, the number of students, school type (TEIP or non-TEIP), and the inspection delegation responsible for the evaluation (north, centre, and south). This information, included in the external evaluation reports, was extracted manually and recorded in a database for subsequent analysis. Additionally, a latent variable was defined to represent the global sense of the external evaluation, calculated as the sum of scores registered in each evaluation domain.

Measurements and Data Analysis

The sample was organised into two groups: the schools evaluated during the school years 2018-2019 and 2019-2020, and the second, the schools evaluated during 2021-2022 and 2022-2023. No school audits were conducted during the 2020-2021 academic year due to the COVID-19 pandemic. Data was analysed with IBM.SPSS Statistics 28.0.

Firstly, data from the external evaluation reports on the inspectorate's appreciation were statistically described using frequencies, means, standard deviations, kurtosis, and skewness. Following Liddell and Kruschke (2018), normality indices were inspected because treating ordinal scores as continuous may produce distortions when the data are highly skewed or kurtotic. For this reason, normality assumptions were evaluated to ensure that the choice of inferential tests was appropriate. Then, we calculated kurtosis and skewness to avoid bias in the interpretation and created histograms of the data distribution for each variable. The normal distribution curve was displayed to determine, centred on the mean, the direction of the data's drift from 2018-2020 to 2021-2023. In this way, corrections to the data distribution can be provided when data assumptions of normality are violated. The variable external evaluation grade, reflecting the structure of the Inspectorate's evaluative framework, was treated as continuous, and data were displayed in histograms to allow comparison of the tendency between (i) sets of schools evaluated in 2018-2020 and the ones evaluated in 2021-2023; (ii) TEIP and non-TEIP schools.

Given the ordinal nature of the domain ratings, the potential non-normal distribution of several variables, and the unequal group sizes, we applied the non-parametric Wilcoxon-Mann-Whitney U test for two unrelated samples to determine whether there are statistically significant differences between the two groups of schools in each inspection domain and in the overall evaluation. A significance level of $\alpha = 0.05$ was adopted. The null hypothesis tested was that the two populations did not differ in the inspectorate's appreciation.

RESULTS

Sample Characterisation

The sample includes 194 schools; 4 (2.1%) were evaluated in 2018-2019, 56 (28.9%) in 2019-2020, 64 (33.0%) in 2021-2022, and 70 (36.1%) in 2022-2023 (**Table 1**). Most schools evaluated by the inspectorates belong to the north (84; 43.3%) and south delegacies (80; 41.2%). The sample included 18 (9.3%) non-grouped schools and 176 (90.7%) clusters of 2 to 23 schools ruled by the same principal, with an average number of 1,514 students per school. The TEIP program encompasses 39 schools.

The leadership and management domain is the most favourable, with 104 (53.6%) appreciations of 'very good' and 65 (33.5%) of 'good' ($\mu = 3.57$). Conversely, the self-evaluation domain showed the lowest appreciation, with 107 (55.2%) 'good' schools, followed by 46 (23.7%) 'very good' schools ($\mu = 3.03$). Excellence reviews are scarce, especially in self-evaluation (1 school; 0.5%), educational service (3 schools; 1.5%), and results domains (6 schools; 3.1%). The leadership and management domain exhibits 11 (5.7%) registers of 'excellent'.

Table 1. Sample descriptive statistics regarding the external evaluation (N=194)

	Self-evaluation	Leadership and Management	Educational Service	Results	Global Evaluation
Mean	3.03	3.57	3.37	3.23	13.21
SD	0.705	0.740	0.625	0.636	2.348
Kurtosis	-0.173	0.806	-0.080	0.502	-0.296
Skewness	-0.133	-0.660	-0.319	0.650	0.557
Frequencies					
Insufficient	2 (1.0%)	2 (1.0%)	0	0	
Sufficient	38 (19.6%)	12 (6.2%)	12 (6.2%)	16 (8.2%)	
Good	107 (55.2%)	65 (33.5%)	101 (52.1%)	124 (63.9%)	
Very Good	46 (23.7%)	104 (53.6%)	78 (40.2%)	48 (24.7%)	
Excellent	1 (.5%)	11 (5.7%)	3 (1.5%)	6 (3.1%)	

'Insufficient' evaluations are restricted to the self-evaluation and leadership and management domains, with 2 (1.0%) frequencies.

Hypothesis Testing

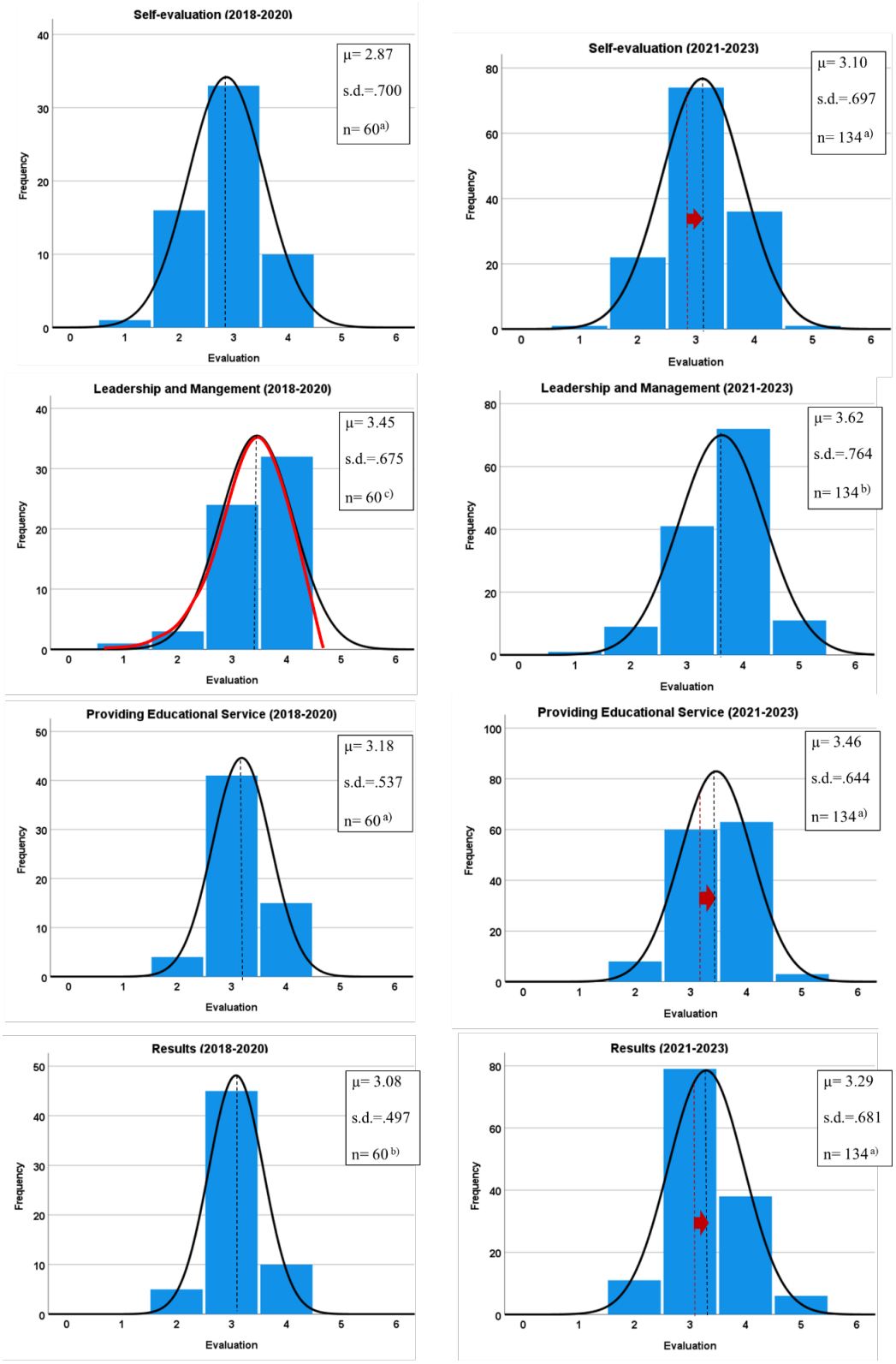
To test whether external evaluation drifts over time and between territories, we compare the schools evaluated in 2018-2020 with those assessed in 2021-2023, as well as those included in the TEIP program with those not covered by it. To determine whether the difference between the external evaluation results for the schools' paired sets is statistically significant, we conduct an inferential analysis using the Wilcoxon-Mann-Whitney *U* test (**Table 2**).

Hypothesis testing of external evaluation differences over time

To test whether external evaluation drifts over time, we compared schools evaluated from 2018 to 2020 and those from 2021 to 2023. **Figure 2** presents the descriptive statistics for both sets of schools. The results reveal that the mean values in all evaluation domains were higher in the 2021-2023 set of schools. The self-evaluation domain increased from 2.87±0.70 to 3.10±0.70; the leadership and management domain increased from 3.45±0.68 to 3.62±0.76; the educational service domain increased from 3.18±0.54 to 3.46±0.64; the results domain increased from 3.08±0.50 to 3.29±0.68.

Table 2. Wilcoxon-Mann-Whitney test results of schools evaluated between 2018 and 2023

	Self-evaluation	Leadership and Management	Educational Service	Results	Global Evaluation
Middle rank					
2018-2020	85.88	89.38	81.61	87.00	80.91
2021-2023	102.71	101.13	104.62	102.20	104.93
U–Mann-Whitney	3322.500	3533.00	3065.000	3390.000	3024.500
Significance	0.032	0.134	0.003	0.040	0.005
Middle rank					
TEIP	100.71	90.51	90.18	84.27	90.67
Non-TEIP	96.69	99.26	99.34	100.83	99.22
U–Mann-Whitney	2897.500	3530.000	3517.000	3390.000	3024.500
Significance	0.658	0.333	0.307	0.053	0.390



Note: a) Kurtosis and skewness between -1 and 1; b) Skewness between -1 and 1; kurtosis between -2 and 2; c) Kurtosis and skewness are not between -1 and 1; red line – distribution correction

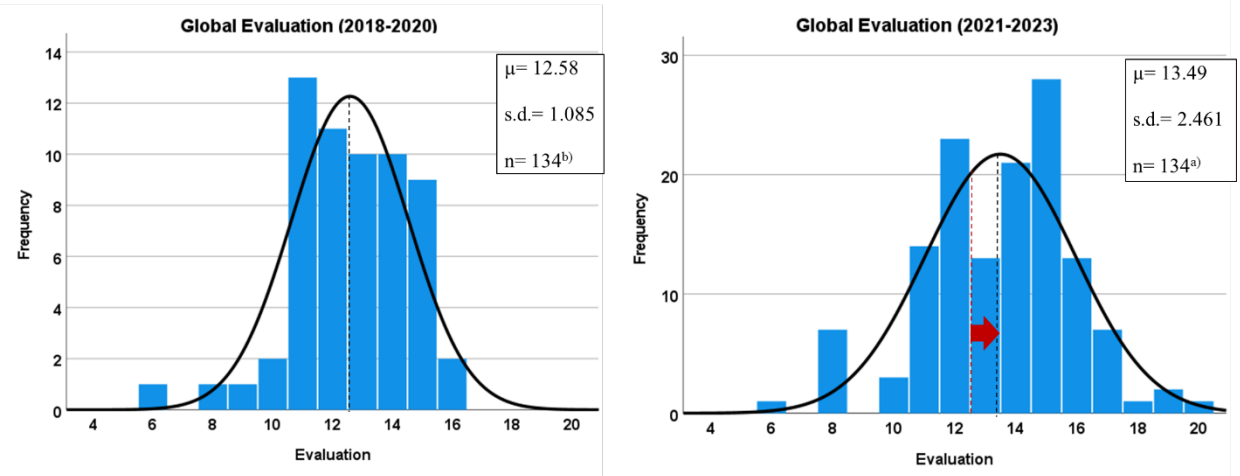
Figure 2. Evolution of the external evaluation results between 2018-2020 and 2021-23

These results led to an increase in the global evaluation from 12.58±1.95 to 13.49±2.46 (Fig. 3). Regarding the self-evaluation domain, the most frequent evaluation was 'good', accounting for around 55% in both sets of schools, even though 'sufficient' scores decreased, and 'very good' scores increased.

In the leadership and management domain, the prevailing evaluation is 'very good', with similar values across both sets of schools (around 53%). However, the frequency of 'good' schools diminished, while the number of 'excellent' schools increased. Regarding educational services, the most frequent evaluation in 2018-2020 was 'good' (68.3%), and in 2021-2023, it was 'very good' (47.0%). Regarding the results domain, 'good' is the prevailing evaluation in both sets of schools. However, it diminished from 75% in 2018-2020 to 59% in 2021-2023, accompanied by an increase of 'very good' and 'excellent' evaluations.

The comparison between the two unrelated sets of schools (**Table 2**) reveals no statistically significant differences between the external evaluations of leadership and management domains over time (U = 3533, p = 0.134). On the other hand, differences regarding the self-evaluation (U = 3322.500, p < 0.05), educational service (U = 3065.000, p < 0.01), and results (U = 3390.000, p < 0.05) support that there are statistically significant differences between the evaluations registered in 2018-2020 and 2021-2023. These differences justify and contribute to the statistically significant difference observed in the global evaluation, with U = 3024.500 (p-value < 0.01) for both sets of schools. In light of this evidence, hypothesis 1 is partially supported. We can infer that the leadership and management domain follows a stabilising evolutionary model (**Figure 1**).

In 2018-2020, the distribution of external evaluation was not normal, considering that the kurtosis and skewness values are out of the cutoff points -1 and 1, but is stabilising around the mean point in 2021-2023 (Fig. 2). The former, with values for asymmetry between -1 and 1 and kurtosis between -2 and +2, is considered acceptable to prove normal univariate distribution (George & Mallery, 2003). On the other hand, regarding the remaining three domains, hypothesis 1 is not supported by data, as the differences between evaluations observed in 2018-2020 and 2021-2023 are statistically significant. It suggests that external evaluation assumes a directional evolutionary model (**Figure 1**) concerning the self-evaluation domain, given the drifting of +0.23 considered the mean point (**Figure 2**) and becoming more favourable; educational service is sliding +0.28 around the mean point and is also becoming more favourable; the results domain drifted +0.21 from the mean point and is also improving. In light of the prevailing results regarding the four domains of external evaluation, the global analysis also rejected hypothesis 1. The evidence gathered documents a drift of +0.91 and more favourable evaluations between the set of schools evaluated in 2018-2020 and 2021-2023 (**Figure 3**).



Note: a) Kurtosis and skewness between -1 and 1; b) Skewness between -1 and 1; kurtosis between -2 and 2.

Figure 3. Evolution of the global external evaluation results between 2018-2020 and 2021-2022

Hypothesis testing of external evaluation differences due to context

To test whether the external evaluation differs by context, TEIP and non-TEIP schools evaluated from 2018 to 2023 were compared. The results in both sets of schools reveal that the mean values are similar concerning the self-evaluation domain (around 3.00). However, the leadership and management domain (TEIP = 3.46±0.790; non-TEIP = 3.59±0.727), the educational service domain (TEIP = 3.26±0.677; non-TEIP = 3.40±0.609), and the results domain (TEIP = 3.05±0.686; non-TEIP = 3.27±0.617) reveal minor differences regarding the context under analysis, with higher values in the non-TEIP set of schools. These slight differences are reflected in the global evaluation, with means of 12.85±2.498 and 13.30±2.309 in TEIP and non-TEIP schools, respectively.

The self-evaluation domain exhibits the lowest evaluation scores in both contexts, with the frequency of 'good' (TEIP = 52.3%; non-TEIP = 56.1%) and 'sufficient' (TEIP = 20.5%; non-TEIP = 19.4%) evaluations. 'Good' is also the most frequent evaluation concerning the results domain (TEIP = 61.5%; non-TEIP = 64.5%), even though 'sufficient' is more representative in TEIP reality, with 17.9% than in non-TEIP schools (5.8%), where 'very good' is the second highest score (26.5%). Educational service registered high frequencies of 'good' in both schools' sets (TEIP = 48.7%; non-TEIP = 52.9.5%), followed by 'very good' evaluations (TEIP = 38.5%; non-TEIP = 40.6%). The leadership and management domain exhibits higher 'very good' frequencies (TEIP = 48.7%; non-TEIP = 54.8%).

The comparison between the two unrelated sets of schools reveals no statistically significant differences between the two contexts in the four evaluation domains (**Table 2**). The Wilcoxon-Mann-Whitney *U* test comparing TEIP and non-TEIP schools yields a *p*-value greater than 0.05, indicating that the null hypothesis cannot be rejected. Then, the data do not support hypothesis 2, and the external evaluation does not discriminate between TEIP and non-TEIP realities.

DISCUSSION

Justice, transparency, accuracy, exemption, and effectiveness are epithets that should characterise an accountability system. External evaluation and inspectorates must overcome the trust deficit with schools (Simeonova et al., 2020) and affirm their role as a force for educational quality and schools' organisational and pedagogical responsiveness. The external evaluation must clearly assert its purpose as a contributor to interactive, constructive, articulated, and intelligent accountability. The idea that "quality" seems to be construed in terms of ambiguity, as compliance with standards (Mufic, 2023) and founded on actions to prepare for the inspection (Six, 2021), demands an intelligent accountability system that asserts itself as a self-improving system. Regarding the pressure signals in the Portuguese educational context, demarcated by a high-stakes inspection system, the empirical evidence gathered points to the alignments discussed below, in which political and socio-economic purposes are transversal.

System Quality

The data clarify that the Portuguese school evaluation system is not static over time. The data suggest that schools, across the self-evaluation, educational services, and results domains, are trending towards more positive evaluations. This piece of evidence can have four possible explanations. First, the external evaluation serves as a resource to improve educational quality, given the trend towards better evaluations observed in the schools evaluated between 2021 and 2023. However, this scenario is shadowed by the fact that penalties regarding a high-stakes system can induce side effects due to evasive behaviour (Thiel & Bellmann, 2017) and pressure, exercised on professionals through leaders (Penninckx, 2017). Then, reasonable doubt arises about whether there is real progress in the quality of the schools. In fact, studies developed in Portugal, concerning the third cycle of external evaluation, point out that self-evaluation in schools needs to evolve to confer higher

centrality to teaching and learning, evaluation cultures must be improved, and self-evaluation must be more integrative and invest in communication to assure reflexivity and involvement in the community (Serra et al., 2024).

The second explanation for the evaluation drift is the ghost of penalties that may induce apparent constructed realities and artefacts, according to Six (2021) and Afonso (2015), stemming from actions taken in preparation for the inspection. This condition may erode trust in the system (Zamir, 2019), exacerbate distortions, and conflict with a self-improving system (Serra et al., 2024).

The third explanation concerns justice in accountability. The schools evaluated at the beginning of the cycle (which tend to receive less positive evaluations) may be harmed, given the direct effect of external evaluation on teachers' and principals' careers. The lack of justice is a significant issue that can lead to further distortion. Transparency regarding the framework and criteria for the schools' evaluation is fundamental to ensure confidence between the schools and the inspectorate by engaging teachers and school leaders (Ehren & Bachmann, 2020). However, accessing factual privileged knowledge produced by the former process of schools' evaluation can provide an advantage in an ongoing cycle. This argument is another element that can contribute to evasive behaviour, compromise trustworthy evaluations, and undermine the core purpose of schools' external evaluation of pursuing quality.

The fourth element that can justify the external evaluation drift is circumstantial, considering the outcomes of Portugal's educational policies. The beginning of the new inspection cycle in 2018, marked by a novel, refined framework and procedures, overlapped with incoming curricular reform aligned with the desideratum announced in Future Education and Skills (OECD, 2018). Schools evaluated in this period underwent organisational and pedagogical transformations, which may have accentuated difficulties during the external evaluation, especially for schools with slower adaptive features. Again, the penalties associated with the external evaluation of schools could lead to injustice in the process and divert attention from school improvement.

According to these four explanations, the corruption of the social process, as explained in Campbell's law (Campbell, 1979), can shadow the accountability process in Portugal and similarly in other high-stakes accountability systems. A distorted system, sustained by pressure and the inevitable circumstances of reality, becomes blind because the external evaluation cannot fully exercise its primary purpose – to promote quality through involvement and implication, self-improvement and continuous improvement, and external support and orientation. High-stakes accountability systems cannot affirm as a provider of quality education because:

> "*When an accountability exercise is riddled with deception, in transparent decision-making, blame games, hidden agendas, or misuse of power on the side of the accountability agent, trust in the accountability system is clearly broken. Only if malpractices of this kind are absent, it seems possible to reap the benefits of accountability and raise the improvement capacity and quality of education systems in a sustainable manner*" (Ehren & Bachmann, 2020, pp. 199-120).

The aspiration is to reject accountability structures built around high-stakes performative measures and appeal for more intelligent accountability (Hutt & Lewis, 2021). The growing idea of external accountability, which supports strong internal accountability, is defensible against others that follow economic features, induce mistrust, pervert transparency, and block school improvement (Cochran-Smith, 2021). For this reason, some approaches describe the evaluator's role as a "critical friend" (Fleischer & Christie, 2009), a partnership in the schools' self-evaluation process, the soul of a quality-focused system, because it is based on proximity and is continuous. However, structural changes and "pedagogical transformation are about a change in cultural assumptions, which entails a slow process of cognitive and emotional modification that has to be supported beyond school walls by concerted social and economic actions" (Mincu, 2022, p. 235). In Portugal, although the external evaluation of schools reflects the globalised education policy, emphasising the

significance of transnational discourses, it must grow as a formative instrument for quality, as its trends are rarely replicated as intended in school practice (Sousa et al., 2021). Ensuring system cohesion in organisational and pedagogical matters requires trust and unity among all educational actors, including politicians, inspectors, principals, middle leaders, and, especially, teachers. Otherwise, the system's coherence will remain weak and stagnant, trapped in its own hypocrisy.

System Standardisation and Legitimisation

An accountability system can lead to standardisation when solutions identified by the system hierarchy as positive in a context are adopted at the expense of gaining internal and social recognition. Fair and legitimate accountability systems will likely endorse coercive and mimetic isomorphism processes and create normative pressures to confirm these standards (Ehren & Visscher, 2006). Isomorphism with institutional rules (Roberts, 2018) can be induced by school inspections, pressures or rewards that lead schools to focus on the performance indicators in the inspection framework standards (Ehren & Visscher, 2006). Isomorphism can, in fact, maintain or even amplify the pseudomorphism effect identified by Serra et al. (2024), which pushes schools to operate within outdated, traditional frameworks due to systemic inconsistencies, given school actors' incomprehension of perspectives arising from mirrored contexts or imposed inspection framework standards. In fact, the drift of external evaluation over time - meaning better-recognised self-evaluation, better educational services provided, and better academic results - may be due to practical improvements introduced in the system or to constructed alignments with the inspection framework. For instance, a qualitative study departing from the external evaluation perspective developed in Portugal attests that collaboration still does not seem to be the norm, even though the role of schools' leadership and management in creating conditions that foster teachers' collaborative work is valued by the inspection (Seabra et al., 2022). Another study regarding external schools' evaluation highlights a shadowed perspective regarding the implementation of supervision practices, which are absent or fragile (Seabra et al., 2021). Additionally, it was observed that schools with higher decisional capital, meaning with practices of mobilisation of knowledge from a solid self-evaluation, exhibit more substantial organisational capital (Serra et al, 2024), even though there are schools that change strictly what is necessary to correspond to inspection suggestions (Seabra et al., 2021). In a high-stakes accountability system, insistent "accountability cultures can produce a certain short-termism in teachers and impede engagement in productive, longer-term, evidence-informed practice" (Brown & Zhang, 2017). A standards-based accountability system may have counterproductive effects on practice as much as a productive impact (Knapp & Feldman, 2012).

Regarding the absence of differences between TEIP and non-TEIP schools, the Portuguese inspectorate framework demonstrates independence and flexibility in responding to the contextual realities of individual schools. Although the logic of legitimisation can also be equated, regarding the TEIP political decision for inclusion, the monitoring conducted by government central services yields results aligned with the schools' progression, with minor discrepancies noted in non-TEIP schools. This aspect is translated into the external evaluation of schools. The slight differences, not statistically significant, between TEIP and non-TEIP schools concerning the domain of the result sustain both the evaluation compliance and the legitimising logic. Achieving impactful, profound changes in school practices is challenging. Sampaio and Leite (2016) found that the impact of external evaluation in TEIP schools is perceived as moderately positive, with higher effects of the leadership and on the strengthening of self-assessment processes, but with less impact on improving student outcomes, their behaviour, the school's relationship with families, and on promoting equity and curricular justice. In Portuguese schools, the difference in evaluations between the leadership and management domain and the remaining domains of schools' external evaluation, which we observed (**Figure 2**), suggests that external evaluation may support surface-level alignment and procedural improvement, but reaching deep cultural transformation is a greater challenge. Indeed, real and sustained changes in school culture, routines, organisational mechanisms, and pedagogical innovation require long-term internal work that inspection

processes alone are not designed to accomplish. This calls for stronger connections between external and self-evaluation and for less hierarchical approaches.

The empirical evidence proves that the external evaluation domain of leadership and management is more stable. This domain is the strongest, and in contrast to the other three external evaluation domains, it has remained essentially unchanged from 2018 to 2023, across both TEIP and non-TEIP schools. This evidence highlights the impact of legitimising the principal, the school's organisation and strategy, and the confidence in the educational system. Inspection added authority and legitimacy to their agendas and deliberations (Cochran-Smith, 2021), thereby enhancing the sustainability of principals' leadership. Portuguese investigation regarding the present cycle of schools' evaluation reveals that despite the 'very good' leadership evaluation, the inspectorate identified fragilities regarding middle leadership and less favourable evaluations concerning results and the schools' educational service (Serra et al., 2023a). Paradoxically, such strong leadership does not achieve the same intensity and results in schools, highlighting system fragilities. This scenario has been discussed in the Portuguese educational context (Barroso, 2022; Castro & Alves, 2013; Serra et al., 2023a), and the literature describes that the principals play a critical role in the process of accountability policy enactment since they frequently establish the conditions under which policy interpretation and enactment would be carried out (Constantinides, 2022).

Limitations

The study model provided three conceptions that guided the data organisation process to identify possible side effects resulting from the schools' external evaluation. However, limitations regarding the exploratory character of the study resulted in more questions regarding (i) the real impact and effectiveness of the external evaluation in promoting schools' improvement cycles and (ii) the existence of staging processes during auditing, translated as a *cat-and-mouse* effect. For clarity, case studies and studies of post-inspection observations are necessary to provide a better framework and to explain the results obtained.

CONCLUSION

This study assembled six main ideas regarding the Portuguese external evaluation process of control and regulation of the quality of the educational system:

1. The schools' external evaluation process behaves like a directional evolutionary model, meaning the evaluation scores drift into higher marks, concerning the self-evaluation, educational services, and results domains.
2. Analysing the Portuguese schools' evaluation process over time suggests the existence of pressure, a common effect in high-stakes accountability systems that can result in crises regarding the purposeful finality of improving educational quality.
3. Over time, data regarding the external evaluation process advises about the risk of standardisation and uniformisation due to top-down inspectorate influence.
4. Processes of the school leaders' legitimisation may emerge from the external evaluation activity, seeming to draw on the logic of staging.
5. Given the high-stakes alignment of the Portuguese schools' evaluation system, side effects in the evaluation process, namely evasive behaviour, apparent and constructed realities, and evaluation distortion, are possible.
6. The framework developed for the third cycle of Portuguese schools' evaluation seems flexible in gathering information from TEIP and non-TEIP schools without discrimination in unfavourable contexts.

On this path, this study contributes new empirical evidence to the literature on educational accountability by demonstrating, through a national dataset of external evaluations, how high-stakes accountability systems can produce upward score drift. Simultaneously, it provides a frame for discussing the risks of standardisation

pressures and symbolic compliance effects over time. By empirically documenting these dynamics in the Portuguese context, the study clarifies mechanisms rarely examined longitudinally and adds comparative value to international research on high-stakes accountability. The findings also illuminate how external evaluation pronounces leadership legitimation processes, offering a more nuanced understanding of how regulation, improvement, and organisational behaviour intersect in real policy settings. As Levantino et al. (2024) defend, this calls for understanding "*the stakes as a continuum in which different types of consequences interact and feed back on one another*" (p. 53).

In conclusion, we do not reject the external accountability in education. Like Cochran-Smith (2021), we defend that accountability "should be reimagined" and follow the paradigm of serving "democratic and humanistic values based on respect for human dignity, equity, diversity, and shared responsibility for a sustainable future" (p.12). An integrated approach sustained in policy visions that favour regulation over control, co-construction over dissociation, and values teachers, contexts, and territories may better serve quality. We argue that politicians should focus on providing and building solid organisational and professional systems in education by following purposeful, intelligent accountability and a culture of whole-system responsibility, rather than high-stakes accountability. In light of these findings, the study identifies several practical recommendations for school leaders, policymakers, and inspectorate services, which are presented in the following section.

PRACTICAL RECOMMENDATIONS

Against the background derived from this study, practical recommendations for schools' governance purposes should be a target of reflection. These proposals derive directly from the empirical patterns observed and seek to strengthen the transparency, fairness, and improvement-orientation of the external evaluation system. In conclusion, notwithstanding the schools' drift toward higher scores in their external evaluations, it cannot be said (for sure) that the quality of teaching and learning has improved, as the risks inherent in a high-stakes inspection system mark the Portuguese system. The idea that solid accountability in a system may automatically support cultures of improvement and leverage quality and inclusion is utopic.

In the face of the piece of evidence gathered, the following practical suggestions may be considered:

Regarding principals, middle leaders, and teachers

- To deconstruct a high-stakes culture of accountability in which the school community may be imbued and embrace intelligent accountability.
- To nurture and invest in a culture of evaluation inscribed in the schools' self-evaluation to sustain action and support decision-making.
- To define strategic cycles of improvement sustained in external evaluation proposals of progress and the schools' self-evaluation.
- To avail the legitimacy generated by the external evaluation regarding the principals' leadership and the school organisational strategy of management to mobilise middle leaders and teachers' agency and operate transformation and inclusion responsiveness.

Regarding policymakers and inspectors:

- To drift from a high-stakes accountability system of regulation to a no- or low-stakes system and flow into intelligent accountability approaches regarding education.
- To eliminate rewards and punishments linked with the schools' external evaluation.
- To invest in providing orientations regarding strengths and improvement areas in external evaluation reports without connotating them with quantitative or qualitative metrics.
- To invest in mechanisms of follow-up and capacity-building in schools that need support to deal with inclusion responsiveness.

- To value and support cooperative or bottom-up approaches and avoid top-down imposition.
- To promote the inspectors' capacity-building to better operate in processes of no- or low-stakes accountability systems.
- To act against and prevent side effects regarding ongoing regulation processes depicted as evasive behaviour and distortion.
- To build an external evaluation mechanism uplifted in trust, transparency, justice, and effectiveness.
- To build an external evaluation process that supports the school's improvement planning, considering the capacity to exercise autonomy according to organisational and pedagogical matters.

Disclosure Statement: Authors confirm no conflicts of interest to report concerning this article.

Data Availability Statement: The data supporting this study's results can be found at http://infoescolas.mec.pt/ and https://www.igec.mec.pt/PgMapa.htm.

REFERENCES

Afonso, A.J. (2015). Do desequilíbrio do pilar da autoavaliação no modelo de avaliação externa: Apontamentos. In E. Faria & R. Perdigão (Eds.), *Textos do seminário avaliação externa das Escolas* (pp. 217-225). Conselho Nacional da Educação. ISBN: 978-972-8360-97-9

Barreira, C., Vaz Rebelo, M. P., Bidarra, M. G., Seabra, F., & Abelha, M. (2023). Cap. III - Satisfação, efeitos e mecanismos de mudança na sequência do 3º ciclo de AEE: perceções de professores e lideranças escolares. In. I. Fialho, et. al. (Eds.), *Avaliação Externa das Escolas. Mecanismos de Mudança nas Escolas e na Inspeção* (pp. 89-116). Editora Humus.

Barroso, J. (2022). *Administração e política educacional: Um percurso de investigação* (Coleção Trajetos de Investigação Educacional) [Ebook]. Instituto de Educação, Universidade de Lisboa.

Brown, C., & Zhang, D. (2017). Accounting for discrepancies in teachers' attitudes towards evidence use and actual instances of evidence use in schools. *Cambridge Journal of Education*, *47*(2), 277–295. https://doi.org/https://doi.org/10.1080/0305764X.2016.1158784

Brunsson, N. (2006) A *Organização da Hipocrisia - Diálogo, Decisão e Acção nas Organizações*. Asa.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*(1), 67–90. https://doi.org/10.1016/0149-7189(79)90048-X

Carvalho, L. M., & Costa, E. (2017). Avaliação externa das escolas em Portugal: atores, conhecimentos, modos de regulação. *Revista Brasileira de Política e Administração da Educação*, *33*(3), 685–705. https://doi.org/10.21573/vol33n32017.79302

Castro, H. de F., & Alves, J. (2013). Avaliação de escolas: o gerenciamento da imagem ao serviço da legitimação. *Revista Portuguesa de Investigação Educacional*, *13*, 49–82. https://doi.org/10.34632/investigacaoeducacional.2013.3389

Cavendish, W., Márquez, A., Roberts, M., Suarez, K., & Lima, W. (2017). Student engagement in high-stakes accountability systems. *Penn GSE Perspectives on Urban Education*, *14*(1), 2–5. http://www.urbanedjournal.org/volume-14-issue-1-fall-2017-15-years-urban-education-special-anniversary-edition-journal/student

Chin, S. Z., Seng, H. Z., & Chan, M. Y. (2019). Doing legitimacy in talk: The production of leader–follower relationship in spiritual consultation interactions. *SAGE Open*, *9*(2). https://doi.org/10.1177/2158244019846695

Cochran-Smith, M. (2021). Rethinking teacher education: The trouble with accountability. *Oxford Review of Education*, *47*(1), 8–24. https://doi.org/10.1080/03054985.2020.1842181

Constantinides, M. (2022). High-stakes accountability policies and local adaptation: exploring how school principals respond to multiple policy demands. *School Leadership and Management*, *42*(2), 170–187. https://doi.org/10.1080/13632434.2021.2016687

Dahler-Larsen, P. (2014). Constitutive effects of performance indicators. *Public Management Review*, *16*(7), 969–986. https://doi.org/10.1080/14719037.2013.770058

Donaldson, G. (2013). Starter Paper on Inspection and Innovation. *Starter Paper on Inspection and Innovation*, 1–8. https://www.nmva.smm.lt/wp-content/uploads/2013/06/SICI-Paper-Bratislava-2013-final-version-24-05-Graham-Donaldson.pdf

Ehren, M., & Bachmann, R. (2020). Accountability to build school and system improvement capacity. In M. Ehren & J. Baxter (Eds.), *Correct: Trust, accountability and capacity in education system reform: Global perspectives in comparative education* (pp. 102-123). Routledge. https://doi.org/10.4324/9780429344855-5

Ehren, M., & Visscher, A. J. (2006). Towards a theory on the impact of school inspections. *British Journal of Educational Studies*, *54*(1), 51–72. https://doi.org/10.1111/j.1467-8527.2006.00333.x

Fleischer, D. N., & Christie, C. A. (2009). Evaluation use: Results from a survey of U.S. American Evaluation Association Members. *American Journal of Evaluation*, *30*(2), 158–175. https://doi.org/https://doi.org/10.1177/109821400833100

Fahey, G. & F. Köster (2019). Means, ends and meaning in accountability for strategic education governance, *OECD Education Working Papers*, No. 204, OECD Publishing, Paris, https://doi.org/10.1787/1d516b5c-en.

Fullan, M., Rincón-Gallardo, S., & Hargreaves, A. (2015). Professional capital as accountability. *Educational Policy Analysis Archives*, *23*(15), 1–18. https://doi.org/10.14507/epaa.v23.1998

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. 11.0 update. Allyn & Bacon.

Hanberger, A., Carlbaum, S., Hult, A., Lindgren, L., & Lundström, U. (2016). School evaluation in Sweden in a local perspective: A synthesis. *Education Inquiry*, *7*(3). https://doi.org/10.3402/edui.v7.30115

Hargreaves, A., & Fullan, M. (2012). *Professional capital: Transforming teaching in every school*. Teachers College Press.

Hargreaves, A. (2019). Teacher collaboration: 30 years of research on its nature, forms, limitations and effects. *Teachers and Teaching: Theory and Practice, 25*(5), 603–621. https://doi.org/10.1080/13540602.2019.1639499.

Heilig, J. V., Brewer, T. J., & Pedraza, J. O. (2018). Examining the myth of accountability, high-stakes testing, and the achievement gap. *Journal of Family Strengths*, *18*(1). https://doi.org/10.58464/2168-670x.1389

Helgøy, I., Homme, A., & Gewirtz, S. (2007). Local autonomy or state control? Exploring the effects of new forms of regulation in education. *European Educational Research Journal*, *6*(3), 198–202. https://doi.org/10.2304/eerj.2007.6.3.198

Hopkins, E., Hendry, H., Garrod, F., McClare, S., Pettit, D., Smith, L., Burrell, H., & Temple, J. (2016). Teachers' views of the impact of school evaluation and external inspection processes. *Improving Schools*, *19*(1), 52–61. https://doi.org/10.1177/1365480215627894

Hutt, M., & Lewis, N. (2021). Ready for reform? Narratives of accountability from teachers and education leaders in Wales. *School Leadership and Management*, *41*(4–5), 470–487. https://doi.org/10.1080/13632434.2021.1942823

Keddie, A. (2015). School autonomy, accountability, and collaboration: A critical review. *Journal of Educational Administration and History*, *47*(1), 1–17. https://doi.org/10.1080/00220620.2015.974146

Knapp, M. S., & Feldman, S. B. (2012). Managing the intersection of internal and external accountability: Challenge for urban school leadership in the United States. *Journal of Educational Administration*, *50*(5), 666–694. https://doi.org/10.1108/09578231211249862

Küçükbere, R. Ö., & Balkar, B. (2021). Teacher accountability for teacher occupational professionalism: The effect of accountability on occupational awareness with the mediating roles of contribution to organization, emotional labor and personal development. *Journal on Efficiency and Responsibility in Education and Science*, *14*(3), 167–179. https://doi.org/10.7160/eriesj.2021.140304

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

Liddell, T.M., & Kruschke, J. K. (2018). Analysing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/https://doi.org/10.1016/j.jesp.2018.08.009

Lillejord, S. (2020). From "unintelligent" to intelligent accountability. *Journal of Educational Change*, *21*(1), 1–18. https://doi.org/10.1007/s10833-020-09379-y

Liu, H., Luo, L., & Tang, W. (2021). Kindergarten teachers' experiences of stress under a high-stakes inspection regime: An exploration in the Chinese context. *International Journal of Educational Research*, *109*(July), 1-10. https://doi.org/10.1016/j.ijer.2021.101850

Matteucci, M., Guglielmi, D., & Lauermann, F. (2017). Teachers' sense of responsibility for educational outcomes and its associations with teachers' instructional approaches and professional well-being. *Social Psychology of Education*, *2*, 275–298. https://doi.org/https://doi.org/10.1007/s11218-017-9369-y

McCrone, T., Coghlan, M., Wade, P., & Rudd, P. (2009). *Evaluation of the impact of Section 5 inspections - Strand 3. Final Report for Ofsted*. Slough: NFER. https://eprints.whiterose.ac.uk/id/eprint/73958/

McNamara, G., & O'Hara, J. (2008). The importance of the concept of self-evaluation in the changing landscape of education policy. *Studies in Educational Evaluation*, *34*(3), 173–179. https://doi.org/10.1016/j.stueduc.2008.08.001

Mincu, M. (2022). Why is school leadership key to transforming education? Structural and cultural assumptions for quality education in diverse contexts. *Prospects*, *52*(3–4), 231–242. https://doi.org/10.1007/s11125-022-09625-6

Miranda, H. M. C. G., Seabra, F., & Pacheco, J. A. (2023). Avaliação externa das escolas, regulação por pares, trabalho colaborativo e qualidade educativa: qual a relação? *Práxis Educativa, 18*, 1–18. https://doi.org/10.5212/PraxEduc.v.18.21863.065

Mouraz, A., Leite, C., & Fernandes, P. (2019). Between external influence and autonomy of schools: Effects of external evaluation of schools. *Paidéia (Ribeirão Preto), 29*, 1-11. http://dx.doi.org/10.1590/1982-4327e2922

Mufic, J. (2023). Discursive effects of "quality" talk during a quality audit in Swedish municipal adult education. *Scandinavian Journal of Educational Research*. https://doi.org/10.1080/00313831.2022.2042844

OECD. (2018). The future of education and skills: Education 2030. *OECD Education Working Papers*, 23. https://one.oecd.org/document/EDU/EDPC(2018)45/ANN2/en/pdf

OECD. (2020). Developing a school evaluation framework to drive school improvement. *OECD Education Policy Perspectives*, *No.26*, OECD Publishing, Paris, https://doi.org/10.1787/60b471de-en.

Ólafsdóttir, B., Jónasson, J. T., Sigurðardóttir, A. K., & Aspelund, T. (2022). The mechanisms by which external school evaluation in Iceland influences internal evaluation and school professionals' practices. *Nordic Journal of Studies in Educational Policy*, *8*(3), 209–224. https://doi.org/10.1080/20020317.2022.2076376

Oliveira, D. S. (2017). *Qualidade da educação em Portugal: O papel da avaliação externa de escolas* [Dissertação de doutoramento, Universidade de Aveiro]. RIA – Repositório Institucional da Universidade de Aveiro.

O'Neill, O. (2013). Intelligent accountability in education. *Oxford Review of Education*, *39*(1), 4–16. https://doi.org/10.1080/03054985.2013.764761

Penninckx, M. (2017). Effects and side effects of school inspections: A general framework. *Studies in Educational Evaluation*, *52*, 1–11. https://doi.org/10.1016/j.stueduc.2016.06.006

Pizmony-Levy, O., & Woolsey, A. (2017). Politics of education and teachers' support for high-stakes teacher accountability policies. *Education Policy Analysis Archives*, *25*(87), 1-26. https://doi.org/10.14507/epaa.25.2892

Portz, J. (2021). "Next-generation" accountability? Evidence from three school districts. *Urban Education*, *56*(8), 1297–1327. https://doi.org/10.1177/0042085917741727

Roberts, J. (2018). Managing only with transparency: The strategic functions of ignorance. *Critical Perspectives on Accounting*, *55*, 53–60. https://doi.org/10.1016/j.cpa.2017.12.004

Sattin-Bajaj, C., & Jennings, J. L. (2020). School counsellors' assessment of the legitimacy of high school choice policy. *Educational Policy*, *34*(1), 21–42. https://doi.org/10.1177/0895904819881774

Schillemans, T., & Bovens, M. (2011). The challenge of multiple accountability: Does redundancy lead to overload? In M. J. Dubnick & H. G. Frederickson (Eds.), *Accountable governance. Problems and promises* (pp. 3–21). Routledge.

Sampaio, M., & Leite, C. (2016). A avaliação externa das escolas e os TEIP na sua relação com a justiça social. *Educação, Sociedade & Culturas*, *47*, 115-136. https://doi.org/10.34626/esc.vi47.190

Seabra, F., Abelha, M., Henriques, S., & Mouraz, A. (2022). Policies and practices of external evaluation of schools: Spaces for teacher collaboration? *Ensaio: Avaliação e Políticas Públicas em Educação*, *30*(116), 664-668. http://dx.doi.org/10.1590/S0104-40362022003003442

Seabra, F., Mouraz, A., Henriques, S., & Abelha, M. (2021). Teacher supervision in educational policy and practice: Perspectives from the External Evaluation of Schools in Portugal. *Education Policy Analysis Archives*, *29*(August - December). https://doi.org/10.14507/epaa.29.6486

Serra, L., Alves, J. M., & Soares, D. R. (2023a). The role of external evaluation control mechanisms and the missing loop of innovation. *Journal of Pedagogical Research*, *7*(5), 156–182. https://doi.org/https://doi.org/10.33902/JPR.202322764

Serra, L., Alves, J. M., & Soares, D. (2023b). Mapping innovation in educational contexts: drivers and barriers. *International Journal of Innovation and Learning*, *35*(1), 74-98. https://doi.org/10.1504/IJIL.2024.135169

Serra, L., Alves, J., & Soares, D. (2024). Pseudomorphosis of schools' system and the fiction of its regulatory processes: A study of educational narratives. *Journal of Pedagogical Research*, *8*(1), 1–27. https://doi.org/10.33902/jpr.202424016

Serra, L., Alves, J. M., & Soares, D. R. (2025). Innovation on the margins of the external evaluation of Portuguese schools. *International Journal of Innovation and Learning*, *37*(1), 60-84. https://doi.org/10.1504/IJIL.2025.143000

Simeonova, R., Parvanova, Y., Brown, M., & Ehren, G. (2020). A continuum of approaches to school inspections: Cases from Europe. *Pedagogy*, *92*(4), 487–507. https://doras.dcu.ie/30486/1/Continuum.pdf

Six, F. (2021). Trust, Accountability and capacity in education system reform: Global perspectives in comparative education. In M. Ehren & J. Baxter (Eds.), *Trust-based accountability in education: The role of intrinsic motivation* (pp. 55–77). Routledge.

Sorm, S., & Gunbayi, I. (2018). School leadership: The exercise of legitimate power in Cambodia. *European Journal of Education Studies*, *4*(1947), 256–284. https://doi.org/10.5281/zenodo.1238513

Sousa, J., & Pacheco, J. A. (2021, September 2-10). *External evaluation of schools in Portugal and the global agenda* [Conference session]. European Conference on Educational Research - ECER 2021 "Education and Society: expectations, prescriptions, reconciliations", Geneva, Italy. https://hdl.handle.net/1822/83627

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), 1-29. https://doi.org/10.14507/epaa.25.2662

Visscher, A. J., & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement*, *14*(3), 321–349. https://doi.org/10.1076/sesi.14.3.321.15842

West, A., & Pennell, H. (2005). Market-oriented reforms and "high stakes" testing: Incentives and consequences*. *Cahiers de La Recherche Sur l'Éducation et Les Savoirs*, *1*, 181–199. https://doi.org/10.4000/cres.1961

Zamir, S. (2019). The polymeric model of school evaluation in the era of accountability. *Quality Assurance in Education*, *27*(4), 401–411. https://doi.org/10.1108/QAE-06-2018-0070