LARGE-SCALE EVALUATION IN LATIN AMERICA
AVALIAÇÃO EM LARGA ESCALA NA AMÉRICA LATINA
EVALUACIÓN EN LARGA ESCALA EN LATINOAMÉRICA

https://doi.org/10.18222/eae.v35.11050

A BRIEF HISTORY OF HIGH-STAKES TESTING AND ITS POSSIBLE FUTURES

MARÍA JESÚS GUTIÉRREZ DOMÍNGUEZ[I]

[I] Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain; *mjgutierrez@uc.cl*

ABSTRACT

Education's historical emphasis on assessments that enhance learning has evolved into a competitive, performance-driven model due to educational reforms and neoliberal influences. This trend has materialised in various forms, and one of them is high-stakes testing. Through a Systematic Literature Review (SLR), the present study develops a global examination of this mechanism, in order to understand its scope and influence on the educational system. The results reveal both adverse and beneficial outcomes of high-stakes testing and, while the literature highlights unintended consequences, it also stresses the importance of its function as an accountability mechanism in the current society, implying positive aspects. Moreover, findings suggest that viable alternatives for large scale assessments extend beyond high-stakes scenarios.

KEYWORDS LARGE SCALE ASSESSMENTS • HIGH-STAKES TESTING • ALTERNATIVE ASSESSMENT • LITERATURE REVIEW.

HOW TO CITE:

Gutiérrez Domínguez, M. J. (2024). A brief history of high-stakes testing and its possible futures. *Estudos em Avaliação Educacional*, *35*, Article e11050. https://doi.org/10.18222/eae.v35.11050

UMA BREVE HISTÓRIA DOS TESTES DE ALTO IMPACTO E SEUS FUTUROS POSSÍVEIS

RESUMO

A ênfase histórica da educação em avaliações que promovem a aprendizagem evoluiu para um modelo competitivo e orientado pelo desempenho impulsionado por reformas educacionais e influências neoliberais. Essa tendência materializou-se de diversas formas, sendo uma delas os testes de alto impacto (*high-stakes testing*). Por meio de uma Revisão Sistemática da Literatura (RSL), o presente estudo desenvolve uma análise global desse mecanismo, buscando compreender seu alcance e impacto no sistema educacional. Os resultados revelam tanto os efeitos adversos quanto os benefícios dos testes de alto impacto Embora a literatura destaque consequências não intencionais, também enfatiza a importância de sua função como um mecanismo de responsabilização na sociedade contemporânea, indicando aspectos positivos. Além disso, os achados sugerem que alternativas viáveis para avaliações em larga escala vão além dos cenários de alto impacto.

PALAVRAS-CHAVE AVALIAÇÃO EM LARGA ESCALA • TESTES DE ALTO IMPACTO • AVALIAÇÃO ALTERNATIVA • REVISÃO DE LITERATURA.

UNA BREVE HISTORIA DE LAS EVALUACIONES DE ALTAS CONSECUENCIAS Y SUS POSIBLES FUTUROS

RESUMEN

El énfasis histórico de la educación en las evaluaciones que promueven el aprendizaje ha evolucionado hacia un modelo competitivo y orientado por el desempeño debido a las reformas educativas y las influencias neoliberales. Esta tendencia se materializó de varias maneras, siendo una las evaluaciones de altas consecuencias (*high-stakes testing*). A través de una Revisión Sistemática de la Literatura (RSL), el presente estudio desarrolla un análisis global de este mecanismo, buscando comprender su alcance e impacto en el sistema educativo. Los resultados revelan resultados tanto adversos como beneficiosos asociados a las evaluaciones de altas consecuencias. Aunque la literatura destaca consecuencias no deseadas, también enfatiza la importancia de su función como mecanismo de responsabilización en la sociedad contemporánea, indicando aspectos positivos. Además, los hallazgos sugieren que las alternativas viables para evaluaciones a gran escala van más allá de los escenarios de altas consecuencias.

PALABRAS CLAVE EVALUACIÓN A GRAN ESCALA • EVALUACIONES DE ALTAS CONSECUENCIAS • EVALUACIÓN ALTERNATIVA • ESTUDIO BIBLIOGRÁFICO.

Received on: MARCH 11, 2024
Approved for publication on: OCTOBER 15, 2024

INTRODUCTION

Throughout the history of education, assessment has been essential for learning. Its main role has historically been as a mechanism for teachers to support learning and enhance students' capabilities and knowledge (Hayward, 2015). Nevertheless, based on new educational reforms since the 1990s, the influence of New Public Management, and neoliberal practices, assessments are now marked by competition and performance mechanisms (Verger, Parcerisa et al., 2019) that are endangering the original essence of evaluation and, consequently, of education. This phenomenon, that Ball (2003, 2012c) denominates 'performativity', narrows and fragments learning (Wyse et al., 2015) and has a built-in over-preoccupation with metrics, measurement, and numbers (Ball, 2015). Thus, it obstructs and limits the possibilities of education, rather than expands and enriches them (Ball, 2012a, 2012b). These trends endanger the learning processes that should occur in classrooms (Madaus & Russell, 2010), as well as influence human beings' essence (Ball, 2017).

Nowadays, what happens internationally in a great number of classrooms depends on a series of factors that are not exactly based on the needs, interests and capabilities of the children that populate those classrooms (Hoyuelos & Cabanellas, 1996). Instead, they are determined by high-stakes testings (HSTs) or large-scale assessments (LSAs) in general and, with them, standardised processes (Ball, 2003, 2017; Madaus & Russell, 2010). Thus, assessments, and the learning that might enhance capabilities and create authentic knowledge, have reduced teachers to focusing on preparing students for the tests (Berliner, 2011). The reason for this is that assessments have become an end in themselves, with a focus on accountability (Hayward, 2015) and rankings (Jones & Ennes, 2018). Likewise, test results are tied to judgements of school and teacher performance and, as long as this dynamic remains unchanged, assessment for learning and all classroom experiences will be determined by it (Hayward, 2015).

Thus, although HSTs fulfil an accountability role that is necessary for society (Bovens et al., 2008) because of the significant data they provide (Schillemans et al., 2013), due to their structure and the stakes involved, they are endangering education, people's subjectivities, learning and the formative meaning of education in the contexts where they take place (Madaus & Russell, 2010; Helfenbein, 2004; Schillemans, 2016; Falabella, 2021). Therefore, the present study will explore the strengths and shortcomings of this mechanism and possible ways to improve the current situation. This rationale and preoccupation translate into a main question, which will be the driving force of this research: How can the future of large-scale assessment serve as both a facilitator for enhanced learning and a robust mechanism for ensuring quality within education systems, and thus improve the current large-scale assessments? This will be done using a Mixed Method Systematic

Literature Review which, on the one hand, will enquire what have been the main consequences of HST through a Review of Reviews and, on the other hand, using a Critical Interpretative Synthesis Review, will explore the best alternatives or possible future for high-stakes testing.

HISTORY OF THE HIGH-STAKES TESTING

The history of high-stakes testing begins with a significant change in governance that occurred in the 1990s, initially in some Western nations (Murphy, 2021) such as the United States, the United Kingdom, and European countries (Levi-Faur, 2012; Lynn, 2012). Over time, it spread to other territories, such as Latin America (Verger, Fontdevila et al., 2019; Rhodes, 1996). This change implied that governments alter their way of governing from a centralised structure in terms of power and control to a decentralised one, consisting of giving both powers and control to new entities such as markets, political agencies or institutions, regional governments, and non-governmental organisations, among others (Levi-Faur, 2012; Rhodes, 2007). Scholars named this phenomenon "New Governance" (Lynn, 2012).

This New Governance (Lynn, 2012) was characterised by "steering at a distance" (Murphy, 2021, p. 53), which was distinguished by regulation, networking, and the creation of standards (Levi-Faur, 2012). Along with this expansion of government and decentralisation came the "problématique" (Levi-Faur, 2012, p. 13). This required that, in order to maintain legitimacy and effectiveness, and preserve the quality of services (Börzel, 2010; Schillemans et al., 2013; Link & Scott, 2010), essentially to demonstrate "Good Governance" (Murphy, 2021, p. 33), many governments had to opt for using mechanisms that the academy has called "managerial technics" (Ackerman, 2004; Hood & Dixon, 2016). These mechanisms or strategies became known as "new public management, or NPM for short" (Murphy, 2021, p. 42).

One of the measures for safeguarding "Good Governance" is high-stakes testing (Nichols & Berliner, 2007), which is a "policy instrument" (Levi-Faur, 2012) that seeks to measure the "learning" or performance of students, teachers, and schools through standardised tests to evaluate education and ensure its quality (Nichols & Berliner, 2007; Muller, 2018). Furthermore, this type of assessment is characterised by the fact that they have "high stakes" (Jones & Ennes, 2018), because they have crucial impacts on educational agents like promoting educators and students, adjusting salaries, and allocating resources (Jones & Ennes, 2018; Gregory & Clarke, 2003).

Naturally, there are multiple causes and factors that influence the consolidation of high-stakes testing. Along with the evolution of governance (Levi-Faur, 2012), there is another aspect involved, the introduction of Western European mass education in the 19th century. This involved the incorporation and expansion

of state-controlled, compulsory general education (Soysal & Strang, 1989) that, according to some authors, was linked to economic development and growth (Zinkina et al., 2016; Westberg et al., 2019). This historical phenomenon included the United States (Beadie, 2019) and then, in the late 19th and early 20th centuries, was introduced into Latin America (Frankema, 2009), Australia, New Zealand and, to a lesser extent, into Asia and Africa (Zinkina et al., 2016).

Mass education was crucial to the formation of national education systems and, thus, to constructing and unifying their national policies (Ramirez & Boli, 1987); it is considered to be "a by-product of industrialization" (Green, 2013, p. 47). Between the 19th and 20th centuries, many changes occurred in the education systems due to mass education, some of which were the unification of the curriculum, regulation of the entry requirements at different levels of the system and, which is more relevant for the present study, education was focused on reproduction, mechanisation, memorisation (Benavot et al., 1991). Subsequently, national assessment measures arose (Green, 2013) and systematisation was essential to demonstrate abilities and distinguish people, giving prominence to meritocracy and competence (Green, 2013).

Additionally, mass education systems entailed the states taking a fundamental role in education: not only in funding but also in regulation and administration which, as mentioned above, gave rise to national curricular and assessment structures (Ramirez & Boli, 1987). Since then, these evaluation systems have evolved and, in the middle of the 19th century, schools started applying standardised tests that were purely memory-based (predominantly oral) (Huddleston & Rockwell, 2015). With the convergence of international influences and historical events, national assessments emerged. Then, with the Second World War, international measurement systems were formalised and spread (Kamens & McNeely, 2010). Although the timing of countries' adoption of these dynamics varies, and some countries have not adopted these mechanisms, LSAs and HSTs have been increasingly expanding their scope, nationally and internationally, and they are shaping each other (Verger, Parcerisa et al., 2019).

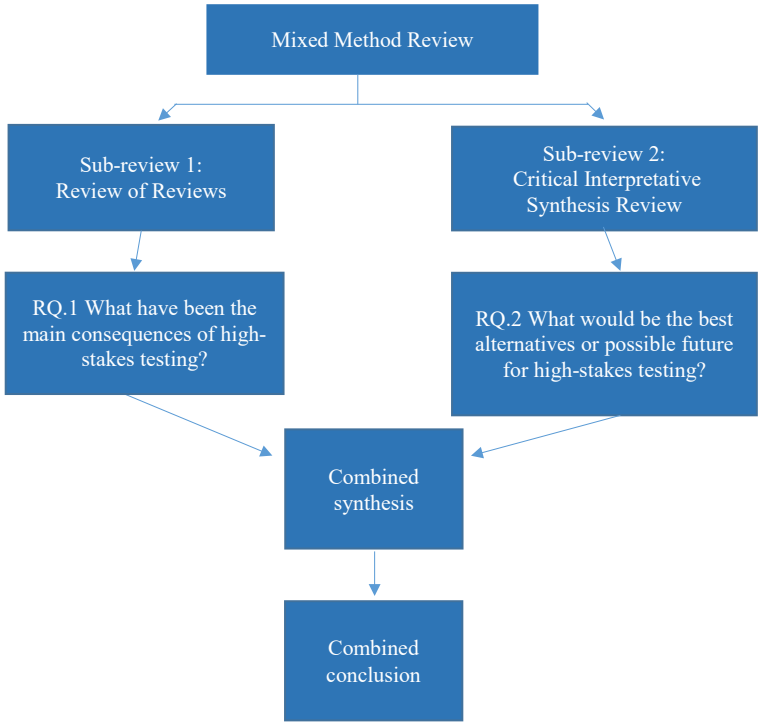
Further on, these large-scale assessments and policy instruments progressively began to take more standardised structures as a result of education reforms and policies (Ball, 2003). Over time, they were imbued with functions that impose significant impacts on individuals and institutions, and so these examinations commenced to have more at stake (Nichols & Berliner, 2007). Thus, performance began to determine the granting of diplomas, access to the next levels of education, teachers' salaries and, more importantly, this new trend created rankings and league tables on which many reputations and billions of pounds are based (Zhao, 2014). Consequently, numbers, statistics, standards, measurements, and performativity became dominant forms in modern society and, with them, high-stakes testing (Ball, 2003; Gregory & Clarke, 2003; Zhao, 2014; Muller, 2018).

METHODOLOGY

The methodology used in the present study is a Systematic Literature Review (SLR), which gives an overview of what is known about a specific topic (Gough et al., 2013). Specifically, this is a Mixed Method SLR because it "has sub-reviews that ask questions about different aspects of an issue" (Gough et al., 2012, p. 6). The first sub-review was a Review of Reviews (Gough et al., 2017), and it attempted to respond to what have been the main consequences of HSTs. For clarity purposes, this research question was assigned the number 1 (RQ.1). The second sub-review was a Critical Interpretative Synthesis Review which summarises data, challenges and questions assumptions that are taken-for-granted (Dixon-Woods et al., 2006, p. 4), and seeks to answer what would be the best alternatives or possible future for HSTs. This research question was assigned the number 2 (RQ.2). Both reviews aim to answer a major question which is, how can the future of large-scale assessment serve as both a facilitator for enhanced learning and a robust mechanism for ensuring quality within education systems and, thus, improve the current LSAs?

This SLR used a mixed methodology because finding patterns in the papers, analysing the data, and creating codes were qualitative; and, associating codes to the patterns and then quantifying them in a graph was quantitative. For a clearer understanding of the methodology, Figure 1, based on Gough et al. (2012), presents a representation of the structure.

FIGURE 1
Structure of the Mixed Method Review conducted in the present study, based on the proposal of Gough et al. (2012)



Source: Author's elaboration (2024).

Stages of the Systematic Literature Review

The stages that were followed to realise this Mixed Method SLR are based on the proposal of the University College of London (UCL) (2023) and Gough et al. (2013). In the first sub-review, 15 papers were selected; in the second sub-review, 25 papers were selected. The commands used for the searches are detailed in appendices A and B. The second sub-review demanded 25 searches, multiple synonyms and the reading of 200 abstracts. Nevertheless, when this second search yielded a scarce number of papers relevant to the question, the Gough et al. (2013) strategy of consulting experts was used. Hence, Professor Clive Dimmock and Dr. Clara Fontdevila provided what, according to their criteria, were relevant articles for the search. More details are in Appendix C.

Inclusion and exclusion criteria used for the selection of the studies are detailed below, in Table 1. It is important to note that the literature review considered exclusively publications written in English.

TABLE 1
Criteria based on Gough et al. (2012) and UCL (2023)

SUB-REVIEW	INCLUSION CRITERIA	EXCLUSION CRITERIA	INCLUSION CRITERIA (IN COMMON)	EXCLUSION CRITERIA (IN COMMON)
1) Review of Reviews	• Only Systematic Literature Reviews were selected	• Every paper not related to the effects or consequences of high-stake testing (or other synonyms)	• Geographic range: international • Language: English • Papers were selected based on their relevance (capacity of response) to their corresponding research questions • Time scope: 1990 to the present day, 2023 • Must contain valid ethical considerations (Petticrew & Roberts, 2006) • Only includes peer-reviewed papers • Only studies focused on high-stake tests on students • Papers could consider effects on children and teachers	• Papers focused on exit examinations • Focused on tertiary or higher education • Focused on impacts on computer assisted testing and evaluation of teachers
2) Critical Interpretative Synthesis Review	• All types of data collection (qualitative and quantitative) • Studies must have a robust methodology and detailed ethical considerations			

Source: Author's elaboration (2024).

Quality appraisal

Firstly, the primary criterion for ensuring quality was the relevance of the selected studies in relation to the research questions (Gough et al., 2013). Secondly, the rigorousness of the methodology (Gough et al., 2013). Thirdly, only peer-revied papers were selected. Fourth and lastly, only papers from journals, and no other documents such as blogs or grey literature were included, to ensure reliability, because blogs and grey literature can be biased, incomplete, and methodologically challenging to evaluate (Hopewell et al., 2005).

Synthesis and data analysis

The present study examined the data, looked for patterns in them, created codes, and values were assigned to those codes. Subsequently, they were organised into charts and interpreted (Gough et al., 2013). To convert the data from qualitative to quantitative, the two reviews were different. Sub-review 1, as it measures consequences and the weight of them, different scores were associated based on how rigurous and robust the papers were. To measure this, the following five quality criteria were selected:

1) Number of studies reviewed.
2) Presents criteria that safeguard quality.
3) Describes the database used.
4) Presents the words used for the search.
5) States the inclusion and exclusion criteria.

Thus, according to the number of criteria that each paper met, a score was assigned to it. These scores, when associated with the consequence codes, would give a final, summative score and, therefore, a weight for each paper. The details of score allocation are given, below:

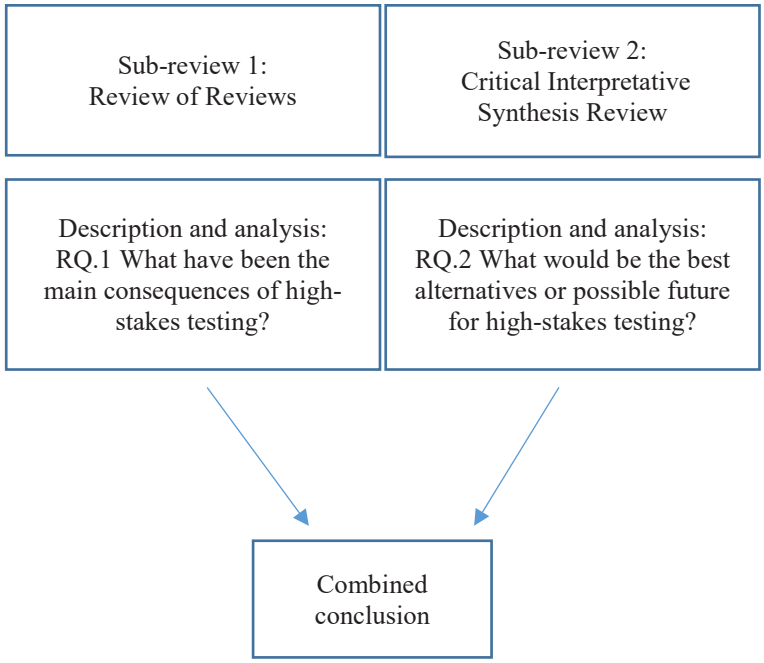
- 3 points: if it meets 4 or more criteria.
- 2 points: if it meets 3 criteria.
- 1 point: if it meets 2 or less of the above criteria.

Likewise, in sub-review 2, as they were the proposals themselves that were important, rather than the robustness of the research, it is the idea itself that matters. Therefore, they all have the same validity.

FINDINGS

Initially, the findings will be explained for each research question, separately. Finally, both research questions will be combined in the discussion and conclusion. This organisation is illustrated in Figure 2.

FIGURE 2
Organisation of the presentation and analysis of results

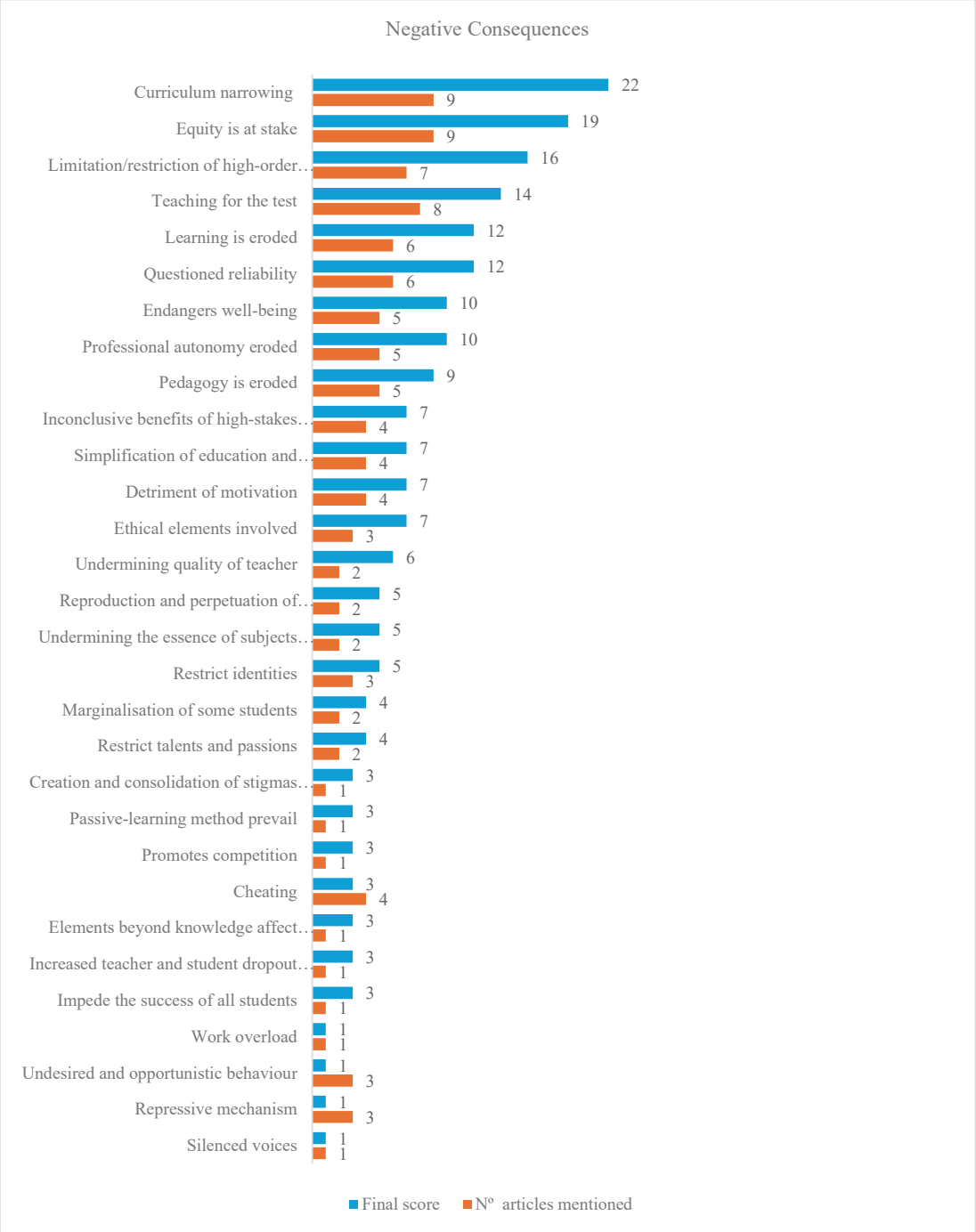


Source: Author's elaboration (2024).

Findings RQ.1 – What have been the main consequences of high-stakes testing?

To identify predominant consequences in the existing literature and to determine those articles that provided substantial support and empirical evidence, all articles were assessed according to their robustness. The following graph, presented in Figure 3, displays the frequency of each consequence mentioned in the papers and, additionally, it assigns a score from 1 to 3 that reflects the level of rigour and robustness. These scores are based on the criteria outlined in the Synthesis and data analysis section.

FIGURE 3
Chart obtained from sub-review 1: Review of Reviews focused on answering RQ.1 – What have been the main consequences of high-stakes testing?



Source: Author's elaboration (2024).

The results of the Review of Reviews suggest that, among the consequences with the highest scores, and those repeated most frequently, were: curriculum narrowing (Acosta et al., 2020; Au, 2009; Sigvardsson, 2017; Boon et al., 2007; Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019; Anderson, 2012; Cimbricz, 2002; Emler et al., 2019; Harlen & Crick, 2003), endangering equity (Acosta et al., 2020; Au, 2009; Boon et al., 2007; Bacon & Pomponio, 2023; Hamilton et al., 2013; Verger,

Parcerisa et al., 2019; Anderson, 2012; Emler et al., 2019; Lee, 2008), limitation or restriction of high-order abilities (Acosta et al., 2020; Au, 2009; Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019; Anderson, 2012, Emler et al., 2019), teachers teaching for the test (Acosta et al., 2020; Au, 2009; Hamilton et al., 2013; Cimbricz, 2002; Ehren et al., 2016; Emler et al., 2019; Nichols, 2007; Harlen & Crick, 2003), learning eroded by the HST (Boon et al., 2007; Hamilton et al., 2013; Anderson, 2012; Cimbricz, 2002; Emler et al., 2019; Nichols, 2007), and the questioned reliability of them (Acosta et al., 2020; Verger, Fontdevila et al., 2019; Hamilton et al., 2013; Cimbricz, 2002; Lee, 2008; Nichols, 2007).

In further detail, the negative consequences after codifications resulted in a total of 30. Similar codes were organised into six themes: (1) the curriculum and what happens in the classroom, (2) how they have influenced teachers, (3) the consistency and reliability of high-stakes testing, (4) school culture, (5) equity, and (6) influence on people's subjectivity. Firstly, the studies reviewed suggest that education has been affected because high-stakes testing has led to curricular narrowing (Acosta et al., 2020; Au, 2009; Boon et al., 2007; Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019; Anderson, 2012; Cimbricz, 2002; Emler et al., 2019; Harlen et al., 2002). Furthermore, this reduction implies a view of the curriculum in its broad spectrum, considering the teaching, content and skills that are developed, as well as the interactions that take place (and do not take place) within the classroom. A concrete example of this is presented by Au (2009), regarding the situation in the United States, "71% of the districts reported cutting at least one subject to increase time spent on reading and math as a direct response to the high-stakes testing mandated under NCLB" (Renter et al., 2006, as cited in Au, 2009, p. 46).

Additionally, several authors stress that high-order abilities such as critical, divergent, and creative thinking do not have priority in the classroom (Acosta et al., 2020; Au, 2009; Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019; Anderson, 2012; Cimbricz, 2002; Emler et al., 2019); thus, reducing teaching (Bacon & Pomponio, 2023; Ehren et al., 2016; Emler et al., 2019) to mere repetition, and memorisation of facts (Au, 2009). Increasingly, teachers have focused on preparing students for the tests (Emler et al., 2019; Nichols, 2007) and, consequently, subjects that are not assessed in the high-stakes tests (Harlen & Crick, 2003), such as art, music, sports, poetry, among others have been diminished in importance (Au, 2007), therefore, culminating in learning detriment (Boon et al., 2007; Hamilton et al., 2013; Anderson, 2012; Cimbricz, 2002, Emler et al., 2019; Nichols, 2007). Acosta et al. (2020, p. 536) highlight a student's opinion, "We're only learning the content of the tests and not what we're supposed to know and go to college".

Secondly, these negative consequences have also had an effect on teachers, their profession and professionalism. On the basis of strong rewards and sanctions,

teachers have fallen into attitudes toward gaming or cheating (Ehren et al., 2016; Emler et al., 2019; Hamilton et al., 2013). HST has led several educators to leave their schools (Boon et al., 2007), has increased their work overload (Ehren et al., 2016), competitiveness (Verger, Parcerisa et al., 2019), endangered teachers' well-being (Anderson, 2012; Cimbricz, 2002; Ehren et al., 2016; Emler et al., 2019; Harlen & Crick, 2003), makes teachers go against their beliefs and values (Ehren et al., 2016), lose their motivation (Hamilton et al., 2013; Emler et al., 2019) erodes the essence of subjects and teaching (Sigvardsson, 2017; Anderson, 2012), among other elements presented in Figure 3. Moreover, along with the harm to teachers, there is the phenomenon of de-professionalisation of teachers, which implies the loss of their professional autonomy (Verger, Fontdevila et al., 2019; Anderson, 2012; Cimbricz, 2002; Ehren et al., 2016; Emler et al., 2019). This has also contributed to the detriment of the quality of teachers (Verger, Parcerisa et al., 2019; Anderson, 2012).

Thirdly, HST fails in the consistency of the information that it provides about learning; 6 of the 15 studies (with a total score of 12) declare that the reliability of the mechanisms is questionable (Cimbricz, 2002; Lee, 2008; Nichols, 2007; Hamilton et al., 2013; Verger, Fontdevila et al., 2019; Acosta et al., 2020); 4 papers with a total of 7 points mention that the studies on benefits are inconclusive (Hamilton et al., 2013; Verger, Parcerisa et al., 2019; Lee, 2008; Nichols, 2007); and 1 study with 3 points, therefore a strong study, declares that performance measurement and scores reflect outcomes that are not directly related to learning and knowledge (Boon et al., 2007). Information is restricted, as it is limited in most cases by multiple-choice questions or closed-ended questions and, in this scenario, HST is not able to provide sufficient information to assess learning and its complexity (Boon et al., 2007; Acosta et al., 2020).

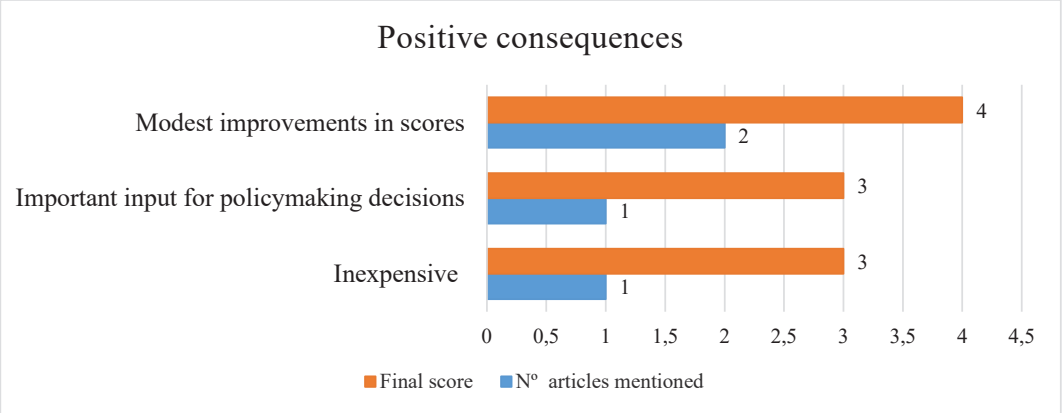
Fourth, studies suggest that HST has generated a competitive culture in schools, and the education system in general (Verger, Fontdevila et al., 2019), where teachers have even turned against some students who perform poorly, marginalising them (Bacon & Pomponio, 2023; Hamilton et al., 2013). As a result, education, teachers, learning, and pedagogy itself, have been affected (Au, 2009; Hamilton et al., 2013; Anderson, 2012; Cimbricz, 2002; Emler et al., 2019).

Fifth, there is equity, which was mentioned in 9 of the total of 15 papers (which is substantial in relation to the other consequences). Those studies indicate that equity is at risk due to HST, because tests tend to perpetuate injustice and inequalities (Emler et al., 2019; Bacon & Pomponio, 2003; Harlen & Crick, 2003) that affect students with disabilities, those who are English learners, and those from disadvantaged communities (Boon et al., 2007). Studies also suggest that HST increases racial and socio-economic disparities (Acosta et al., 2020; Emler et al., 2019). Emler et al. (2019, p. 589) provide more details noting that,

> LSAs have consistently revealed large gaps in scores among different groups of students. The gaps are primarily a result of socioeconomic and racial inequality and other factors that schools and teachers cannot control. . . . In other words, the efforts to close the achievement gap have widened the opportunity gap, creating more inequity and injustice.

Sixth, one of the most striking consequences of HST is the erosion of the human complexity of students, teachers, and education itself (Emler et al., 2019; Acosta et al, 2020; Bacon & Pomponio, 2023; Ehren et al., 2016). Due to standardisation and homogenisation (Emler et al., 2019), the narrowing of the curriculum and teaching (Acosta et al., 2020; Boon et al., 2007), the prevalence of passive learning methods (Anderson, 2012), and the restriction to low-level cognitive skills such as memorisation and repetition (Au, 2009), subjectivity, diversity, and individualisation have been put at risk. This is mentioned in 4 studies with a total score of 7 (Au, 2009; Emler et al., 2019; Harlen & Crick, 2003; Hamilton et al., 2013). Additionally, other studies declare that high-stakes testing limits the possibility for all students to succeed (Acosta et al., 2020) and that it restricts the development of some talents and passions (Emler et al., 2019; Harlen & Crick, 2003).

FIGURE 4
Chart obtained from sub-review 1: Review of Reviews focused on answering RQ.1.2 – What have been the main consequences of high-stakes testing?



Source: Author's elaboration (2024).

The three positive consequences identified in Figure 4 were found by reviewing these 15 papers. These consequences were drawn from 4 different texts, 3 of which are three-point texts, and therefore are strong, rigorous, and robust studies. This demonstrates that these implications are legitimate, important and should be considered (Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019; Lee, 2008; Anderson, 2012).

As shown in Figure 4, these studies present positive consequences of HST related to structural and functional elements. They are centred mainly on political elements, policy-making and financial implications (Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019) which aim for transparency and good governance, and indicate that high-stakes testing responds to a need, especially in terms of accountability (Anderson, 2012). In addition, it is relevant to note that the current large-scale assessment mechanisms are inexpensive in comparison to other options, and therefore attractive for policymaking in different countries (Verger, Fontdevila et al., 2019).

Another consequence is a moderate increase in scores present in two papers (Lee, 2008; Anderson, 2012). This is controversial because it raises the question of how those scores were increased or what those scores mean. These increases in scores could be related to one of the phenomena presented in the negative consequences such as cheating, teaching to the test, the narrowing of the curriculum, among others (Hamilton et al., 2013; Boon et al., 2007). There are critical elements that weaken the argument that favours the improvement in performance, for example, the questions on its reliability (6 papers, score of 12) (Verger, Fontdevila et al., 2019), and the inconclusive beneficial effects that several papers (4 with a score of 7) yielded (Verger, Parcerisa et al., 2019; Hamilton et al., 2013). The reality is that there are contradictory results.

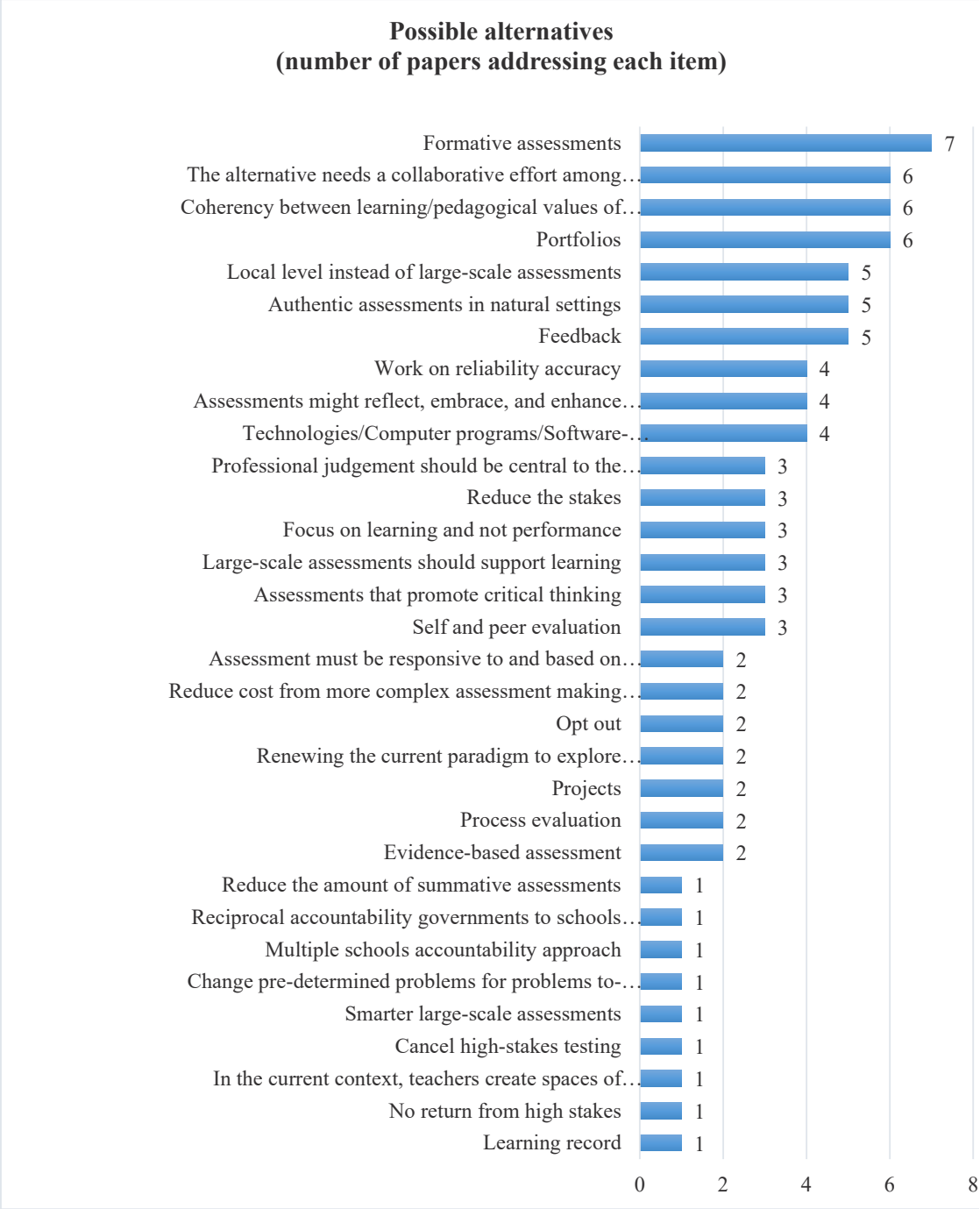
In summary, HST has strengths that should be considered when thinking about future possibilities for these types of evaluation. Although HST is able to respond to a valid need for objectivity, these mechanisms have substantial, unintended consequences with negative effects on education, teachers, students, their learning and individualities, and compromise ethical and equity factors. Therefore, for the future, it is necessary to reconcile strengths with the improvement of weaknesses.

Findings RQ.2 – What would be the best alternatives or possible future for high-stakes testing?

The following compilation represents a sample of the extensive body of literature on the subject. The results presented are based on the review of the 25 selected articles.

FIGURE 5

Chart obtained from sub-review 2: Critical Interpretative Synthesis Review focused on answering RQ.1.3 – What would be the best alternatives or possible future for high-stakes testing?



Source: Author's elaboration (2024).

Most of the results shown in Figure 5 provide ideas and possibilities for improvements to the current large-scale assessment methods. However, two papers suggest abandoning them completely, opting either for the exit of students and schools (Wang, 2017; Ashadi et al., 2022), or cancelling them (Ashadi et al., 2022). These options, although extreme, present a position, but in relation to what is

mentioned previously, they fulfil a current need and have strengths that are relevant when thinking about possible futures.

Conversely, the rest of the reviewed literature recognises the value of these mechanisms for gathering and providing information about learning, the quality of education, responding to the need of public educational institutions to be accountable, and allowing for obtaining input about processes that enables in-depth discussions about processes occurring in the classroom (Dorn, 1998; Chudowsky & Pellegrino, 2003). Likewise, HST functions as a student selection and allocation tool at crucial educational junctures, such as progressing from primary to secondary school, advancing from school to higher education, and making decisions on resource allocation (Suto & Oates, 2021). They also indicate the urgent need to improve, rethink current HST and reconsider the current paradigm (Chudowsky & Pellegrino, 2003; Volante, 2007). Otherwise, the negative consequences and their weaknesses will continue to have a negative impact on learning and perpetuate inequalities (Volante, 2007; Lingard, 2009; Syverson, 2011).

The results of the present review were organised into three clusters. First, concrete alternatives that can provide simpler solutions; second, systemic changes that involve more complex and deeper alternatives; and third, suggestions regarding how the processes of developing new alternatives to high-stakes testing may be conducted.

Firstly, the authors reviewed suggest: the incorporation of feedback (Cato & Walker, 2022; Chudowsky & Pellegrino, 2003; Brown et al., 2014; Zimmerman & Dibenedetto, 2008; Beyond Test Scores Project [BTS Project] & National Education Policy Center [NEPC], 2023), consideration and combination of evaluation in authentic and spontaneous contexts (Syverson, 2011; Roberson, 2011; Brown et al., 2014; Volante, 2007; Lingard, 2009), process evaluation (Behizadeh & Lynch, 2017; Zimmerman & Dibenedetto, 2008), self- and peer-evaluation (Açıkalın, 2014; Chudowsky & Pellegrino, 2003; Roberson, 2011), projects (Açıkalın, 2014; BTS Project & NEPC, 2023), more emphasis on formative assessments to reduce the prominence of summary assessments (Açıkalın, 2014; Chudowsky & Pellegrino, 2003; Roberson, 2011; Brown et al., 2014; Gillanders et al., 2021; Zimmerman & Dibenedetto, 2008; Hutchinson & Hayward, 2005; Hayward et al., 2004; Hayward & Spencer, 2010; BTS Project & NEPC, 2023), foster critical thinking and high-level skills (Ab Kadir, 2017; Roberson, 2011; Brown et al., 2014), focus on learning rather than performance (Volante, 2007; Lingard, 2009; BTS Project & NEPC, 2023), and allow spaces for non-pre-determined forms of evaluation (Beghetto, 2019).

More concretely, some authors recommend: portfolios (Syverson, 2011; Açıkalın, 2014; Chudowsky & Pellegrino, 2003; Behizadeh & Lynch, 2017; Herman & Winters, 1994; BTS Project & NEPC, 2023) and smarter large-scale assessments

through the use of technologies, computer programmes, software and artificial intelligence that allow information to be gathered in a formative manner, providing feedback and many of the aforementioned elements (Beghetto, 2019; Chudowsky & Pellegrino, 2003; Behizadeh & Lynch, 2017).

Secondly, regarding the systemic and complex alternatives, the studies propose that policy should aim for coherence between curriculum and assessment, and that both should be aligned with learning and pedagogical purposes (Ab Kadir, 2017; Dorn, 1998; Chudowsky & Pellegrino, 2003; Volante, 2007; Zimmerman & Dibenedetto, 2008; BTS Project & NEPC, 2023). Using the same logic, several papers point out that it is necessary for these mechanisms to promote and address the complexity and comprehensiveness of human beings and education (Roberson, 2011; Volante, 2007; Hayward & Spencer, 2010; BTS Project & NEPC, 2023). Furthermore, the literature suggested that large-scale assessments should be more focused at local levels (Dorn, 1998; Gillanders et al., 2021; Moss, 2022; Ashadi et al., 2022; Volante, 2007; BTS Project & NEPC, 2023).

Several studies emphasised the importance of devolving responsibility and trust in the professional role of teachers, returning their professional autonomy (Hutchinson & Hayward, 2005; Hayward & Spencer, 2010; Lingard, 2009). Additionally, authors mentioned that the biggest problem with HST is the stakes, and therefore, regardless of what decision is made based on improvements or alternatives, the form of assessment must reduce the stakes (Behizadeh & Lynch, 2017; Hooge et al., 2012; BTS Project & NEPC, 2023).

Thirdly, regarding how to undertake the process of transformation, there are two crucial elements. The first is the need for a collaborative approach, involving all stakeholders and specialists (Chudowsky & Pellegrino, 2003; Volante, 2007; Behizadeh & Lynch, 2017; Hutchinson & Hayward, 2005; Hooge et al., 2012; BTS Project & NEPC, 2023). The second is the fact of rethinking and embracing new paradigms, different from current thinking beyond the boundaries of the system (Chudowsky & Pellegrino, 2003; Volante, 2007).

Lastly, Syverson (2011, p. 4) perfectly illustrates the focus needed for the possible future of HST, saying that "standardized testing has focused on standardizing the content of what is assessed, rather than standardizing the architecture in which diverse kinds of evidence of learning can be collected, organized, understood, and evaluated".

DISCUSSION

The present study found that high-stakes testing does indeed have substantively harmful, negative consequences (Verger, Fontdevila et al., 2019). Yet, at the same

time, it provides an opportunity and is a necessary form of transparency and democracy, fundamental to today's world (Schillemans et al., 2013; Bovens, 2010). Moreover, as Hayward (2015, p. 38) points out, evaluation has to be "as, for and of" learning; therefore, stakeholders involved in education have the opportunity to reverse and improve the situation and transform these mechanisms to the benefit of learning and education. However, this requires two elements to be considered: collaboration that enables people to see what cannot be seen individually and to find solutions that have not yet been considered (Chudowsky & Pellegrino, 2003; Volante, 2007; Hutchinson & Hayward, 2005; Hooge et al., 2012; BTS Project & NEPC, 2023); and, that policy makers permit themselves to analyse new structures and paradigms (Volante, 2007; Chudowsky & Pellegrino, 2003).

Expanding on the previous idea, concrete examples appear in the second sub-review, demonstrating that new movements and possibilities have emerged. In the United States, some individuals and schools have opted out of HST systems; they have created the "FairTest" movement (Syverson, 2011). Likewise, in Scotland, a formative assessment project has also started and has been accompanied by the government in partnership with the academy and schools (Hutchinson & Hayward, 2005; Hayward et al., 2004; Hayward & Spencer, 2010). Using the same logic, the solutions and options presented in the studies yield simple, concrete and meaningful alternatives, such as, for example, the lowering of stakes (Hooge et al., 2012; BTS Project & NEPC, 2023). This is a small step, but it would have significant effects on classrooms, schools and learning, possibly reducing some of the unintended consequences.

On the other hand, addressing only superficial aspects, rather than tackling the root cause of a problem, will only maintain the problem. For instance, focusing solely on specific content or skills as a remedy. In the case of PISA, which has incorporated creativity into its assessments, this could lead countries and schools to adopt unethical behaviours, teaching to the test, and narrowing curricula to prioritise this new skill (Beghetto, 2019). Therefore, achieving meaningful and profound improvements requires addressing problems at their foundation (Beghetto, 2019).

Similarly, another highly realistic solution is to implement different forms of assessment (Hooge et al., 2012), complementing the summative with the formative, and facilitating it through group assessments by territory (Volante, 2007; BTS Project & NEPC, 2023). In this way, it is not overly expensive, and the high amount of effort required would be done in small stages. In addition, the incorporation of technologies in the day-to-day classroom allows the evaluation of processes and giving of feedback from a more ludic and pedagogical perspective, which is also a reasonable and effective alternative (Behizadeh & Lynch, 2017; Beghetto, 2019). Lastly, there is the proposal of thinking about assessment forms without pre-determined

ideas, with open-ended possibilities of questions or exercises, allowing students to determine their own ideas, without markers having a pre-determined idea of a correct answer (Beghetto, 2019).

Finally, it is relevant to consider that education not only shapes the system itself but also influences society and, consequently, humanity: "what students should know and what students do not know are all highly controlled by the examinations" (Emler et al., 2019, p. 281). The consequences of this situation become more apparent when individuals are not aware of their knowledge gaps. Thus, individuals are not even able to realise that there are things they do not know, but this is a topic for further investigation. However, as long as high-stakes assessments persist as they are now, the content taught in classrooms will be limited to what is measured. Everything outside that spectrum will remain unknown or unexplored (Emler et al., 2019).

CONCLUSION

Overall, the present study initially addressed the history of HST, related concepts, and phenomena in order to facilitate understanding and enable the reader to contextualise. Then, the methodology delved into the two sub-reviews that constituted the general SLR, which facilitated an in-depth enquiry. The subsequent step involved presenting the obtained results, followed by the discussion, implications, and recommendations.

Hence, it is necessary to return to the main research question, presented at the beginning of this study: how can the future of large-scale assessment serve as both a facilitator for enhanced learning and a robust mechanism for ensuring quality within education systems? The answer is composed of a few ideas. First, there is a need for collaborative partnerships that consider students and policymakers. From there, it is necessary to recognise what works and what does not, and think about how to improve it. While questioning and understanding underlying power dynamics is fundamental, it is necessary to move toward mechanisms that empower everyone.

The literature already contains some of this completed work, including the main consequences and ways to improve. However, it is now time to create, to think beyond the boundaries, and to put together the pieces of the puzzle that were presented in the present study. Thus, it is time to build mechanisms that utilize formative and summative assessments (Brown et al., 2014) in authentic settings (Syverson, 2011), that enhance high-order skills (Hayward & Spencer, 2010), that give space for the to-be-determined problems (Beghetto, 2019), that generate local solutions and use technologies (Behizadeh & Lynch, 2017). Likewise, reduction of the stakes should be reconsidered (Hooge et al., 2012). The introduction of resources

such as portfolios (Chudowsky & Pellegrino, 2003), projects (BTS Project & NEPC, 2023), peer and self-assessment (Açıkalın, 2014), the incorporation of software that measures processes and provides feedback (Behizadeh & Lynch, 2017), among many other options that are detailed in Figure 5 should be explored. However, it is necessary to highlight that, whatever the assessment is, it must respond to the context in which it is immersed and be a means of promoting and becoming part of the learning process. In essence, a balance can be established between enhanced learning and a robust mechanism for ensuring quality within education systems.

In conclusion, the changes require that governments take the initiative and begin to think collaboratively outside the boundaries to create new LSA mechanisms. These mechanisms should allow for creative, complex futures in which education preserves its primary purpose, learning, and allows students to succeed and flourish according to each of their interests and talents (Nussbaum, 2011; Emler et al., 2019), that expand the possibilities of education and pursue equity rather than perpetuate inequalities (Zhao, 2014).

REFERENCES

Ab Kadir, M. A. (2017). Engendering a culture of thinking in a culture of performativity: The challenge of mediating tensions in the Singaporean educational system. *Cambridge Journal of Education*, *47*(2), 227-246. https://doi.org/10.1080/0305764X.2016.1148115

Ackerman, J. (2004). Co-governance for accountability: Beyond "exit" and "voice". *World Development, 32*(3), 447-463. https://doi.org/10.1016/j.worlddev.2003.06.015

Açıkalın, M. (2014). Future of social studies education in Turkey. *Journal of International Social Studies, 4*(1), 93-102. https://www.iajiss.org/index.php/iajiss/article/view/130

Acosta, S., Garza, T., Hsu, H.-Y., Goodson, P., PadrÛn, Y., Goltz, H. H., & Johnston, A. (2020). The accountability culture: A systematic review of high-stakes testing and english learners in the United States during no child left behind. *Educational Psychology Review, 32*, 327-352. https://doi.org/10.1007/s10648-019-09511-2

Anderson, K. J. (2012). Science education and test-based accountability: Reviewing their relationship and exploring implications for future policy. *Science Education*, *96*(1), 104-129. https://doi.org/10.1002/sce.20464

Ashadi, A., Margana, M., Mukminatun, S., & Utami, A. B. (2022). High stakes testing cancellation and its impact on EFL teaching and learning: Lessons from Indonesia. *International Journal of Language Education*, *6*(4), 397-411. https://doi.org/10.26858/ijole.v6i4.34743

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267. https://doi.org/10.3102/0013189X07306523

Au, W. (2009). Social studies, social justice: W(h)ither the social studies in high-stakes testing? *Teacher Education Quarterly*, *36*(1), 43-58.

Bacon, J., & Pomponio, E. (2023). A call for radical over reductionist approaches to 'inclusive' reform in neoliberal times: An analysis of position statements in the United States. *International Journal of Inclusive Education*, *27*(3), 354-375. https://doi.org/10.1080/13603116.2020.1858978

Ball, S. J. (2003). The teacher's soul and the terrors of performativity. *Journal of Education Policy, 18*(2), 215-228. https://doi.org/10.1080/0268093022000043065

Ball, S. J. (2012a). *Foucault, power, and education*. Routledge.

Ball, S. J. (2012b). *Global Education Inc.: New policy networks and the neoliberal imaginary*. Taylor & Francis Group.

Ball, S. J. (2012c). Performativity, commodification and commitment: An I-Spy guide to the neoliberal university. *British Journal of Educational Studies, 60*(1), 17-28. https://doi.org/10.1080/00071005.2011.650940

Ball, S. J. (2015). Education, governance and the tyranny of numbers. *Journal of Education Policy, 30*(3), 299-301. https://doi.org/10.1080/02680939.2015.1013271

Ball, S. J. (2017). *Foucault as educator*. Springer.

Beadie, N. (2019). "Hidden" governance or counterfactual case? The US failure to pass a national education act, 1870-1940. In J. Westberg, L. Boser, & I. Brühwiler (Eds.), *School acts and the rise of mass schooling: Education policy in the long nineteenth century* (pp. 325-348). Springer International Publishing.

Beghetto, R. A. (2019). Large-scale assessments, personalized learning, and creativity: Paradoxes and possibilities. *ECNU Review of Education, 2*(3), 311-327. https://doi.org/10.1177/2096531119878963

Behizadeh, N., & Lynch, T. L. (2017). Righting technologies: How large-scale assessment can foster a more equitable education system. *Berkeley Review of Education, 7*(1), 25-47. https://doi.org/10.5070/B87130877

Benavot, A., Cha, Y.-K., Kamens, D., Meyer, J. W., & Wong, S.-Y. (1991). Knowledge for the masses: World models and national curricula, 1920-1986. *American Sociological Review, 56*(1), 85-100. https://doi.org/10.2307/2095675

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education, 41*(3), 287-302. https://doi.org/10.1080/0305764X.2011.607151

Beyond Test Scores Project (BTS Project), & National Education Policy Center (NEPC). (2023, Spring). *Educational accountability 3.0: Beyond ESSA*. BTS Project; NEPC.

Boon, R., Voltz, D., Lawson, C., & Baskette, M. (2007). The impact of high-stakes testing for individuals with disabilities: A review synthesis. *Journal of the American Academy of Special Education Professionals*, 54-67. https://files.eric.ed.gov/fulltext/EJ1140225.pdf

Börzel, T. A. (2010). *Governance with/out government: False promises or flawed premises?* [Working Paper No. 23]. SFB Governance. https://ciaotest.cc.columbia.edu/wps/sfb/0018726/f_0018726_16022.pdf

Bovens, M. (2010). Two concepts of accountability: Accountability as a virtue and as a mechanism. *West European Politics, 33*(5), 946-967. https://doi.org/10.1080/01402382.2010.486119

Bovens, M., Schillemans, T., & Hart, P. T. (2008). Does public accountability work? An assessment tool. *Public Administration, 86*(1), 225-242. https://doi.org/10.1111/j.1467-9299.2008.00716.x

Brown, N. J., Afflerbach, P. P., & Croninger, R. G. (2014). Assessment of critical-analytic thinking. *Educational Psychology Review, 26*(4), 543-560. https://doi.org/10.1007/s10648-014-9280-4

Cato, H., & Walker, K. (2022). The influences of teacher knowledge on qualitative writing assessment. *Journal of Language and Literacy Education, 18*(2), 1-21.

Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice, 42*(1), 75-83.

Cimbricz, S. (2002). State-mandated testing and teachers' beliefs and practice. *Education Policy Analysis Archives, 10*(2), 1-21. https://doi.org/10.14507/epaa.v10n2.2002

Dixon-Woods, M., Cavers, D., Agarwal, S., Annandale, E., Arthur, A., Harvey, J., Hsu, R., Katbamna, S., Olsen, R., Smith, L., Riley, R., & Sutton, A. J. (2006). Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology, 6*(1), Article 35. https://doi.org/10.1186/1471-2288-6-35

Domina, T., Penner, A., & Penner, E. (2017). Categorical inequality: Schools as sorting machines. *Annual Review of Sociology, 43*, 311-330. https://doi.org/10.1146/annurev-soc-060116-053354

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, 6(1), 1-33. https://doi.org/10.14507/epaa.v6n1.1998

Ehren, M. C., Jones, K., & Perryman, J. (2016). Side effects of school inspection; motivations and contexts for strategic responses. In M. C. M. Ehren (Ed.), *Methods and modalities of effective school inspections* (pp. 87-109). Springer. https://doi.org/10.1007/978-3-319-31003-9_5

Emler, T. E., Zhao, Y., Deng, J., Yin, D., & Wang, Y. (2019). Side effects of large-scale assessments in education. *ECNU Review of Education, 2*(3), 279-296. https://doi.org/10.1177/2096531119878964

Falabella, A. (2021). The seduction of *hyper-surveillance*: Standards, testing, and accountability. *Educational Administration Quarterly, 57*(1), 113-142. https://doi.org/10.1177/0013161X20912299

Frankema, E. (2009). The expansion of mass education in twentieth century Latin America: A global comparative perspective. *Revista de Historia Económica – Journal of Iberian and Latin American Economic History, 27*(3), 359-396. https://doi.org/10.1017/S0212610900000811

Gillanders, C., Iruka, I. U., Bagwell, C., & Adejumo, T. (2021). Parents' perceptions of a K-3 formative assessment. *School Community Journal, 31*(2), 239-266. https://files.eric.ed.gov/fulltext/EJ1323055.pdf

Gough, D., Oliver, S., & Thomas, J. (2013). *Learning from research: Systematic reviews for informing policy decisions*. Alliance for Useful.

Gough, D., Oliver, S., & Thomas, J. (2017). Introducing systematic reviews. In D. Gough, S. Oliver, & J. Thomas (Eds.), *An introduction to systematic reviews* (2nd ed., pp. 1-17). Sage.

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

Green, A. (2013). *Education and state formation: Europe, East Asia and the USA*. Palgrave Macmillan.

Gregory, K., & Clarke, M. (2003). High-stakes assessment in England and Singapore. *Theory into Practice, 42*(1), 66-74. https://doi.org/10.1207/s15430421tip4201_9

Hamilton, L. S., Schwartz, H. L., Stecher, B. M., & Steele, J. L. (2013). Improving accountability through expanded measures of performance. *Journal of Educational Administration, 51*(4), 453-475. https://doi.org/10.1108/09578231311325659

Harlen, W., & Crick, R. D. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice, 10*(2), 169-207. https://doi.org/10.1080/0969594032000121270

Harlen, W., Crick, R. D., Broadfoot, P., Daugherty, R., Gardner, J., James, M., & Stobart, G. (2002). *A systematic review of the impact of summative assessment and tests on students' motivation for learning* (EPPI-Centre Review). EPPI-Centre.

Hayward, L. (2015). Assessment is learning: The preposition vanishes. *Assessment in Education: Principles, Policy & Practice, 22*(1), 27-43. https://doi.org/10.1080/0969594X.2014.984656

Hayward, L., Priestley, M., & Young, M. (2004). Ruffling the calm of the ocean floor: Merging practice, policy and research in assessment in Scotland. *Oxford Review of Education, 30*(3), 397-415. https://doi.org/10.1080/0305498042000260502

Hayward, L., & Spencer, E. (2010). The complexities of change: Formative assessment in Scotland. *Curriculum Journal*, *21*(2), 161-177. https://doi.org/10.1080/09585176.2010.480827

Helfenbein, R. J. (2004). New times, new stakes: Moments of transit, accountability, and classroom practice. *Review of Education, Pedagogy, and Cultural Studies*, *26*(2-3), 91-109. https://doi.org/10.1080/10714410490480368

Herman, J. L., & Winters, L. (1994). Portfolio research: A slim collection. *Educational Leadership, 52*(2), 48-55.

Hood, C., & Dixon, R. (2016). Not what it said on the tin? Reflections on three decades of UK public management reform. *Financial Accountability & Management*, *32*(4), 409-428. https://doi.org/10.1111/faam.12095

Hooge, E., Burns, T., & Wilkoszewski, H. (2012). *Looking beyond the numbers: Stakeholders and multiple school accountability* [Working Papers No. 85]. OECD Education. https://dx.doi.org/10.1787/5k91dl7ct6q6-en

Hopewell, S., Clarke, M., & Mallett, S. (2005). Grey literature and systematic reviews. In H. R. Rothstein, A. J. Sutton, & M. Borenstein, *Publication bias in meta-analysis: Prevention, Assessment and Adjustments* (pp. 49-72). Wiley. https://doi.org/10.1002/0470870168.ch4

Hoyuelos, A., & Cabanellas, I. (1996). Malaguzzi y el valor de lo cotidiano [Presentación de trabajo]. *Congreso de Pamplona*, Pamplona, España. https://www.waece.org/biblioteca/pdfs/d091.pdf

Huddleston, A. P., & Rockwell, E. C. (2015). Assessment for the masses: A historical critique of high-stakes testing in reading. *Texas Journal of Literacy Education*, *3*(1), 38-49.

Hutchinson, C., & Hayward, L. (2005). The journey so far: Assessment for learning in Scotland. *Curriculum Journal*, *16*(2), 225-248. https://doi.org/10.1080/09585170500136184

Jones, M. G., & Ennes, M. (2018). High-stakes testing. *Oxford Bibliographies*. https://doi.org/10.1093/obo/9780199756810-0200

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review, 54*(1), 5-25. https://doi.org/10.1086/648471

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Educational Research, 78*(3), 608-644. http://www.jstor.org/stable/40071139

Levi-Faur, D. (2012). From "big government" to "big governance"? In D. Levi-Faur (Ed.), *The Oxford Handbook of Governance* (pp. 3-18). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199560530.013.0001

Lingard, B. (2009). Testing times: The need for new intelligent accountabilities for schooling. *QTU Professional Magazine, 24*, 13-19.

Link, A. N., & Scott, J. T. (2010). Historical perspectives on public accountability. In A. N. Link, & J. T. Scott, *Public goods, public gains: Calculating the social benefits of public R & D* (pp. 20-26). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199729685.003.0003

Lynn, L. E. (2012). The many faces of governance: Adaptation? Transformation? Both? Neither? In D. Levi-Faur (Ed.), *The Oxford Handbook of Governance* (pp. 49-64). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199560530.013.0004

Madaus, G., & Russell, M. (2010). Paradoxes of high-stakes testing. *Journal of Education, 190*(1-2), 21-30. https://doi.org/10.1177/0022057410190001-205

Moss, G. (2022). Researching the prospects for change that COVID disruption has brought to high stakes testing and accountability systems. *Education Policy Analysis Archives, 30*(139), 1-24. https://doi.org/10.14507/epaa.30.6320

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton University Press.

Murphy, M. (2021). *Social theory: A new introduction*. Palgrave Macmillan.

Nichols, S. L. (2007). High-stakes testing. *Journal of Applied School Psychology, 23*(2), 47-64. https://doi.org/10.1300/J370v23n02_04

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

Nussbaum, M. C. (2011). *Creating capabilities: The human development approach*. Harvard University Press.

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

Ramirez, F. O., & Boli, J. (1987). The political construction of mass schooling: European origins and worldwide institutionalization. *Sociology of Education, 60*(1), 2-17. https://doi.org/10.2307/2112615

Rhodes, R. A. W. (1996). The new governance: Governing without government. *Political Studies, 44*(4), 652-667. https://doi.org/10.1111/j.1467-9248.1996.tb01747.x

Rhodes, R. A. W. (2007). Understanding governance: Ten years on. *Organization Studies, 28*(8), 1243-1264. https://doi.org/10.1111/j.1467-9248.1996.tb01747.x

Roberson, S. (2011). Defying the default culture and creating a culture of possibility. *Education, 131*(4), 885-904.

Schillemans, T. (2016). Calibrating public sector accountability: Translating experimental findings to public sector accountability. *Public Management Review, 18*(9), 1400-1420. https://doi.org/10.1080/14719037.2015.1112423

Schillemans, T., Van Twist, M., & Vanhommerig, I. (2013). Innovations in accountability: Learning through interactive, dynamic, and citizen-initiated forms of accountability. *Public Performance & Management Review, 36*(3), 407-435. https://doi.org/10.2753/PMR1530-9576360302

Sigvardsson, A. (2017). Teaching poetry reading in secondary education: Findings from a systematic literature review. *Scandinavian Journal of Educational Research, 61*(5), 584-599. https://doi.org/10.1080/00313831.2016.1172503

Soysal, Y. N., & Strang, D. (1989). Construction of the first mass education systems in nineteenth-century Europe. *Sociology of Education, 62*(4), 277-288. https://doi.org/10.2307/2112831

Suto, I., & Oates, T. (2021). *High-stakes testing after basic secondary education: How and why is it done in high-performing education systems?* (Research report). *Cambridge Assessment.*

Syverson, M. A. (2011). Social justice and evidence-based assessment with the learning record. In P. Kriese, & R. E. Osborne, *Social justice, poverty and race* (pp. 93-102). Brill.

University College of London (UCL). (2023). Systematic reviews: Stages in a systematic review. *UCL*. Retrieved May 10, 2023 from https://library-guides.ucl.ac.uk/systematic-reviews/stages

Verger, A., Fontdevila, C., & Parcerisa, L. (2019). 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. (2019). 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

Volante, L. (2007). Educational quality and accountability in Ontario: Past, present, and future. *Canadian Journal of Educational Administration and Policy*, (58), 1-21.

Wang, Y. (2017). The social networks and paradoxes of the opt-out movement amid the Common Core State Standards implementation. *Education Policy Analysis Archives*, *25*(34), 1-27. https://doi.org/10.14507/epaa.25.2757

Westberg, J., Boser, L., & Brühwiler, I. (2019). The history of school acts. In J. Westberg, L. Boser, & I. Brühwiler (Eds.), *School acts and the rise of mass schooling: Education Policy in the long nineteenth century* (pp. 1-15). Springer. https://doi.org/10.1007/978-3-030-13570-6_1

Wyse, D., Hayward, L., & Pandya, J. Z. (Ed.). (2015). *The Sage handbook of curriculum, pedagogy and assessment*. Sage.

Zhao, Y. (2014). *Who's afraid of the big bad dragon? Why China has the best (and worst) education system in the world*. John Wiley & Sons.

Zimmerman, B. J., & Dibenedetto, M. K. (2008). Mastery learning and assessment: Implications for students and teachers in an era of high-stakes testing. *Psychology in the Schools*, *45*(3), 206-216.

Zinkina, J., Korotayev, A., & Andreev, A. (2016). Mass primary education in the nineteenth century. In L. E. Grinin, I. V. Ilyin, P. Herrmann, & A. V. Korotayev, *Globalistics and globalization studies: Global transformations and global future* (pp. 63-70). 'Uchitel' Publishing House.

APPENDICES

Appendix A – Keywords sub-review 1

- S1 – [high stakes testing OR high stakes testing in schools]
- S2 – systematic review OR systematic literature review OR literature review
- S3 – S1 AND S2: [high stakes testing OR high stakes testing in schools] AND [systematic review OR systematic literature review OR literature review]

Plus the filters: Scholarly (peer-reviewed) journals.

Appendix B – Keywords sub-review 2

Final search: **S21** AND **S22**

- **S21** contained: **S19** AND **S20**
- **S19** contained the following terms:

  AB [CLASSROOM environment OR EDUCATIONAL innovations OR EDUCATION & state OR FORMATIVE evaluation OR LEARNING strategies OR EDUCATIONAL programs OR evaluation OR STUDENTS OR rating of OR MIDDLE school education OR PRESCHOOL education OR ELEMENTARY

education OR SCOTLAND OR POLITICAL science VALUATION OR CURRICULUM planning OR EDUCATIONAL change OR FORMATIVE evaluation OR EDUCATIONAL programs OR Evaluation SCOTLAND FORMATIVE evaluation OR TEACHER development; SECONDARY education; ELEMENTARY education OR SCOTLAND OR ORGANIZATIONAL change OR EDUCATIONAL innovations OR RESEARCH OR EDUCATIONAL tests & measurements OR SECONDARY education OR ELEMENTARY education OR EDUCATIONAL evaluation OR RESEARCH OR TEACHERS EDUCATIONAL standards OR SCHOOL environment research OR STUDENT activism OR PROTEST movements] OR AB alternative education OR AB alternative assessment OR AB alternative evaluation AND high stake testings OR large scale assessment AND resistance OR change OR new forms OR new possibilities OR [new possibilities in education evaluation OR large scale assessment]

- **S20** contained the following terms:

  AB high stakes testing OR AB large scale assessment AND AB assessment OR AB standardized testing OR AB standardized test OR AB performativity

- **S22** contained the following terms:

  EDUCATIONAL innovations OR formative evaluation OR EDUCATIONAL change OR assessment for learning

Appendix C – Details of experts consulted in the Sistematic Literature Review

Expert number 1: Clive Dimmock

Recommended Louise Hayward:

1. Hutchinson, C., & Hayward, L. (2005). The journey so far: Assessment for learning in Scotland. *Curriculum Journal, 16*(2), 225-248. https://doi.org/10.1080/09585170500136184
2. Hayward, L., Priestley, M., & Young, M. (2004). Ruffling the calm of the ocean floor: Merging practice, policy and research in assessment in Scotland. *Oxford Review of Education*, *30*(3), 397-415. https://doi.org/10.1080/0305498042000260502
3. Hayward, L., & Spencer, E. (2010). The complexities of change: Formative assessment in Scotland. *Curriculum Journal, 21*(2), 161-177. https://doi.org/10.1080/09585176.2010.480827

Expert number 2: Clara Fontdevila

Recommended the following texts and authors:

1. Hooge, E., Burns, T., & Wilkoszewski, H. (2012). *Looking beyond the numbers: Stakeholders and multiple school accountability* [Working Papers No. 85]. OECD Education. https://dx.doi.org/10.1787/5k91dl7ct6q6-en
2. Lingard, B. (2009). Testing times: The need for new intelligent accountabilities for schooling. *QTU Professional Magazine, 24*, 13-19.
3. Beyond Test Scores Project (BTS Project), & National Education Policy Center (NEPC). (2023, Spring). *Educational accountability 3.0: Beyond ESSA*. BTS Project; NEPC.
4. Domina, T., Penner, A., & Penner, E. (2017). Categorical inequality: Schools as sorting machines. *Annual Review of Sociology, 43*, 311-330. https://doi.org/10.1146/annurev-soc-060116-053354