Performance Funding: Exam Results, Stakes and Washback in Danish Schools

Bukh, Per Nikolaj; Christensen, Karina Skovvang; Poulsen, Morten Lund

Published in:
Sage Open

DOI (link to publication from Publisher):
10.1177/21582440221082100

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Original Research

SAGE Open
January-March 2022: 1–13

DOI: 10.1177/21582440221082100
journals.sagepub.com/home/sgo
SAGE

Performance Funding: Exam Results, Stakes, and Washback in Danish Schools

Per Nikolaj Bukh[1], Karina Skovvang Christensen[2], and Morten Lund Poulsen[2]

Abstract

High-stakes testing is meant to create a positive washback effect on student learning. Performance funding can raise stakes. However, it is not often used, and its washback is uncertain. The purpose of this paper is to examine performance-funding programs based on students' exam results. We study principals' perceptions and interpretations of how this influenced stakes and washback effects of the exit exams. For that purpose, we selected four schools based on theoretical sampling criteria. The empirical data comprise semi-structured interviews with management over the 2-year program and documents describing the performance-funding program. The findings indicate that implementing performance funding increases stakes and has washback effects, but that stakes depend partly on the principal's choices. Although the consequences were unintended, the program and its effects were mostly perceived as positive. The paper shows how unintended consequences call for careful consideration of the pros and cons of accountability systems when high-stakes test-based funding mechanisms are introduced.

Keywords

educational policy, incentives, performance funding, school management, washback effects, teaching to the test

Introduction

The compulsory basic education in Denmark consists of 10 years of primary and lower secondary school education. Most children attend the *Folkeskole*, which is managed by the municipalities. As in the other Nordic countries, new public management-inspired policies have been introduced that make the public education system more focused on accountability and control (Imsen et al., 2017; Myhre, 2021). Following mediocre results in the PISA (Program for International Student Assessment) tests, several policy initiatives were taken to improve the quality of the *Folkeskole* (Ratner, 2020), including mandatory national testing in 2010 (Andreasen et al., 2015; Beuchert & Nandrup, 2018).

In 2016, the government concluded that some student groups still graduated without satisfactory Danish language and mathematics skills. Consequently, the Danish Ministry of Education introduced a policy initiative, the Financing to Enhance Academically Weak Students program (FEAWS). It introduced a performance-funding model (Dougherty et al., 2016) rewarding selected schools with a yearly bonus of 5% to 8% of the annual budget if they reached specific improvement targets for the number of students achieving a threshold exam result.

Performance funding is often used in higher education (e.g., Blankenberg & Phillips, 2016; Ortagus et al., 2020; Umbricht et al., 2017), making university funding partly dependent on output measures, for example, graduation and retention rates. Performance funding in basic education, where school funding is partially dependent on students' achievements, is less widespread. We have identified only a single study (Al-Samarrai et al., 2018) of funding linked directly to students' test results in basic education.

Much research has focused on washback effects, that is, the influence of testing on teaching and learning (Alderson & Wall, 1993, p. 120; Cheng, 1999). Such effects can be perceived and analyzed as positive when they result in "good" teaching practices (Holm & Kousholt, 2019, p. 921) or negative if they adversely affect teaching and learning outcomes. Washback's effects depend on many factors (Alderson & Wall, 1993). Studies have confirmed that high-stakes testing has washback effects on teaching and learning (Akpinar & Cakildere, 2013; Alderson & Wall, 1993; Cheng, 1999; Luxia, 2005). Further, high-stakes testing can have

[1]Aalborg University, Denmark
[2]Aarhus University, Denmark

Corresponding Author:
Per Nikolaj Bukh, Department of Economics and Management, Aalborg University, Fibigerstræde 11, Aalborg 9100, Denmark.
Email: pnb@pnbukh.com

unintended consequences (e.g., Hill et al., 2016; Kelley & Protsik, 1997; Lee & Medina, 2018; Paletta et al., 2020) as negative washback.

Principals play an important role in policy implementation (e.g., Laughlin et al., 1994; Shaked & Schechter, 2017; Urick & Bowers, 2014) and thus the washback of test-based policies. Placed between potentially conflicting external demands and internal, professional goals for teachers (Spillane et al., 2002), principals mediate a policy's influence on teaching (Koyama, 2014; Lambersky, 2016; Shirrell, 2016). Principals' prior experience and strategic choices are important to how they make sense of current situations (Ganon-Shilon & Schechter, 2019) with different possible perceptions of accountability demands (Shirrell, 2016) as well as funding schemes (Pouncey et al., 2013) and their alignment with internal goals. Because of such differences, principals can enact the same policy differently (Shaked & Schechter, 2017, p. 22). Policy implementation is more likely to intensify when aligned with tangible incentives (Spillane et al., 2002, p. 738), that is, when stakes are raised. As stated above, high-stakes testing can lead to negative washback. Nonetheless, Spillane and Kenney (2012, p. 550) suggest that focusing on such adverse unintended consequences of policy implementation can underplay the good faith efforts by principals attempting to accomplish positive washback.

We expect schools' stakes to increase when funding is linked to exam results. However, school management's perceptions and interpretations of performance-funding programs likely influence mediation, implementation, and washback effects. Whether and how performance funding based on students' exit exams influences exam-related stakes to introduce washback effects has yet to be studied. This paper addresses whether performance funding based on exam results has similar consequences as when exam stakes increase through other means, and it aims to answer the following two research questions:

- How do principals perceive performance funding based on exam results?
- How do principals perceive stakes change when performance funding is introduced?
- What are the adverse washback effects, if any, from performance funding based on exit exams?

The paper presents a qualitative study of four Danish public schools participating in the FEAWS program. We interviewed school management twice based on semi-structured interview guides to understand managements' interpretation of the program and changes to educational practices.

This paper contributes to the extant literature by examining whether performance funding based on what was prior low-stakes exams has the same consequences as high-stakes standardized tests and exams. Previous research on high-stakes testing in basic education (Amrein-Beardsley, 2009; Brill et al., 2018; Figlio & Ladd, 2015) and performance funding within higher education (Blankenberg & Phillips, 2016; Ortagus et al., 2020; Umbricht et al., 2017) show both intended and unintended consequences. We contribute by examining whether performance funding within basic education has similar consequences. Furthermore, we explore what behavior is initiated and how principals rationalize it. Thus, this paper contributes to the literature on school management as mediators (Koyama, 2014; Lambersky, 2016; Shirrell, 2016) by demonstrating how principals perceive, interpret, and enact performance funding.

Theoretical Framework

Performance funding is resource allocation based on results (Dougherty et al., 2016; Herbst, 2007: chapter 4). It is mostly associated with tertiary education financing (e.g., Blankenberg & Phillips, 2016; Ortagus et al., 2020; Umbricht et al., 2017), where university funding partly depends on output measures such as graduation rates, retention rates, or graduates' average pay. Such an approach aligns with the focus on output often present in new public management-inspired policies (Myhre, 2021). Although relevant education performance measures include student achievements based on tests, we have identified only one study of funding linked directly to students' basic education test results (Al-Samarrai et al., 2018); this found positive effects on student results on standardized tests.

Washback Effects

Washback effects (Alderson & Wall, 1993), that is, the notion that testing and its structure influence teaching and learning, are commonplace in the education literature (Biggs, 1995; Cheng, 1999). Washback can be perceived as either positive or negative, depending on whether it results in "good" teaching practices (Holm & Kousholt, 2019, p. 921) or undesirable teaching and learning outcomes.

However, washback is a complex phenomenon, and it has been debated whether it is determined by the test itself or factors beyond it. Alderson and Wall (1993, p. 121) suggested that "[t]ests that have important consequences will have washback," whereas "[t]ests that do not have important consequences will have no washback." Numerous studies confirm that high-stakes testing influences teaching and learning (Cheng, 1999). Linking school funding to test scores creates a high-stakes testing situation (Au, 2007) because the results have substantial consequences.

Generally, heightening stakes related to test scores seems to positively affect student outcomes, especially in mathematics (Figlio & Ladd, 2015). However, testing is not a complete measure of student learning or ability (Jennings & Bearak, 2014). Much research (e.g., Amrein-Beardsley, 2009; Brill et al., 2018; Deming et al., 2016; De Wolf & Janssens, 2007) has examined negative washback effects, for example, how "[a]ccountability pressures faced by teachers

and leaders may lead well-intentioned educators to engage in strategic reporting and operational practices to increase test scores, graduation rates, and other indicators of student success" (Edwards & Mindrila, 2019, p. 3). In such cases of negative washback, increases to measured performance can come at a cost to actual performance.

Effects of High-Stakes Testing

Although the categories are relatively fuzzy, negative washback is often identified as different unintended consequences (Amrein-Beardsley, 2009; Paletta et al., 2020) conceptualized as teaching to the test, narrowing the curriculum, and focusing on marginal students, for example, by Spillane and Kenney (2012, p. 542), if under different names. However, such washback cannot be presumed negative because it depends on practice and whether its consequences benefit students.

Teaching to the test represents teaching what is tested rather than what is relevant for learning (Popham, 2001). It displaces other, more relevant activities and inflates test scores (Jennings & Bearak, 2014, p. 382). Teaching to the test practices include, for instance, "students spend[ing] hours memorizing facts, learning test-taking strategies, bubbling score sheets accurately, eliminating unlikely distractor responses, [and] making educated guesses" (Amrein-Beardsley, 2009, p. 3). However, if students become more comfortable in test-taking situations, teaching to the test can enable students to more correctly demonstrate knowledge and skills (Jennings & Bearak, 2014).

Narrowing the curriculum (Amrein-Beardsley, 2009) represents an increased focus on testing at the expense of non-tested parts of the curriculum. Some argue that it is a consequence of teaching to the test (Au, 2007; Brill et al., 2018) because it affects the curriculum. However, as Amrein-Beardsley (2009) demonstrates, curriculum narrowing has broader implications when parts of the curriculum can be skipped for specific students and when the effort is directed toward particular subjects or topics within them. This is, for instance, the case when music teachers engage in test preparations rather than their respective non-tested subjects (Booher-Jennings, 2005, p. 242).

Focusing on marginal students (Burgess et al., 2005; Deming et al., 2016) represents a strategic response to a system that makes some students more valuable for the school. Such students get more attention than others solely due to tests' structure because they are at "the cusp of proficiency" (Golann, 2015, p. 104), that is, close to the testing system's proficiency threshold, and thus have a relatively heavy influence on evaluations.

Principals as Policy Mediators

Principals such as school management, administrators, and leaders are central to school performance (Chua & Mosha, 2015; Kalkan et al., 2020). They allocate resources, build organizational capacity, and promote change (Paletta et al., 2020), influencing teachers' perception, understanding, emotional commitment, and motivation (Lambersky, 2016; Spillane et al., 2002). Principals affect teaching indirectly through these influences on teachers, and they indirectly affect some of the types of washback, for example, teaching to the test. They affect others directly, for example, by allocating resources between subjects, classes, and students.

School management finds itself at the intersection between external accountability demands and the school community (Spillane et al., 2002). In this position, studies have found principals actively negotiating between policy and local practice (Shaked & Schechter, 2017) to selectively prioritize and perform the parts of policies that they deem the most important (Gawlik, 2015; Koyama, 2014). Educational professionals' norms can conflict with accountability policies (Englund et al., 2019; Myhre, 2021), and, in turn, principals prioritizing compliance with such policies (Spillane & Kenney, 2012, p. 549). Although compliance with external demands can prevent school-level punishments or terminations, teachers can respond by making a principal's life miserable (Spillane et al., 2002, p. 747).

Because of schools' different situations and diverging experiences and already made strategic choices, principals can differ in their interpretation, prioritization, and consequently, implementation of policies (Ganon-Shilon & Schechter, 2019; Shirrell, 2016; Spillane et al., 2002). The implementation can differ in intensity, content, and focus (Shaked & Schechter, 2017, p. 22). Solely implementing changes with low to no intensity, for example, to insulate teachers (Laughlin et al., 1994) by paying lip service to formal demands (Ganon-Shilon & Schechter, 2019, p. 238), becomes less likely as the tangible rewards and punishments aligned with implementation (Spillane et al., 2002, p. 738; Spillane & Kenney, 2012), that is, stakes, increase.

The literature shows that principals manage and mediate policy implementation and that the mediation is affected by stakes. Furthermore, it finds that skilful leaders can reconcile policy demands and teachers' norms (Gawlik, 2015; Koyama, 2014; Spillane & Kenney, 2012), but also that the results of high-stakes testing can come in the form of negative washback (e.g., Amrein-Beardsley, 2009; Paletta et al., 2020; Spillane & Kenney, 2012). School managements' mediation of policy has been studied in different contexts, such as the introduction of budgets (Laughlin et al., 1994), the implementation of formalized, mandatory evaluation practices (Paletta et al., 2020), and often in extensive reforms encompassing many aspects of school life (Ganon-Shilon & Schechter, 2019; Gawlik, 2015; Koyama, 2014; Shaked & Schechter, 2017; Spillane & Kenney, 2012). Shaked and Schechter (2017, p. 20) nonetheless argued that the literature on how principals enact policy is still meagre. As far as performance funding based on student exit exams is concerned, its mediation and enactment by principals are yet to be studied.

The Incentive Program

Basic Education in Denmark

The Danish constitution entitles all children aged 6 to 16 years to free education in public schools. In 2018, approximately 77% of this age group were enrolled in 1 of 1,082 public schools averaging 500 students per school. The rest attended various forms of private schools partly financed by the government. The school system is organized such that the Danish Ministry of Education provides the overall objectives and directions for educational goals and management. At the same time, the municipalities are responsible for operating the schools. In practice, the municipalities have high degrees of freedom concerning organizing and funding of the schools, entailing significant differences between municipalities.

Danish educational policy has since World War II been "influenced by progressive pedagogical ideas, which were also reflected in a critical attitude toward grading and examinations" (Andreasen, 2019, p. 138). However, as in other Nordic countries, it has become more focused on accountability, control, and incentives, especially in the new millennium (Imsen et al., 2017)

Denmark participates in PISA's international student assessments, and mandatory national testing was introduced in 2010 (Andreasen et al., 2015; Beuchert & Nandrup, 2018), primarily for development purposes. Besides this, ninth class exit exams provide the only systematic testing of students. The exams hold some stakes to students by partly determining secondary education possibilities. Still, because schools are not rewarded or sanctioned based on tests and exam results, school-level accountability is based on low-powered incentives compared with Anglo-Saxon school systems (Andersen & Nielsen, 2020).

The exit exams consist of mandatory exams in Danish (reading, writing, spelling, and oral), math (calculus and problem-solving), English (oral), and physics/chemistry (oral). The exams are nationwide and decided at the central administrative level but are not considered standardized tests reflecting course objectives and general skills valued by teachers, as Beuchert and Nandrup (2018) emphasize.

The Danish marking scale has seven possible marks and is comparable to the European Credit Transfer and Accumulation System. The scale has five passing marks: 2 (E), 4 (D), 7 (C), 10 (B), and 12 (A) and 2 failing marks: −3 (Fx) and 00 (F). Enrolment in vocational schools requires a GPA of 2, whereas enrolment in the upper secondary school requires a GPA of 5.

Financing to Enhance Academically Weak Students

In 2016, the Danish government decided to "strengthen the efforts of schools for the most disadvantaged students" (Danish Government, 2016). They introduced the FEAWS as a financial incentive program for schools with many students failing to achieve satisfactory levels of academic ability (Danish Ministry of Education, 2017, 2018). The program operationalized *academically weak students* as the relative number of students not achieving a GPA above 4 after the mandatory ninth class examinations in Danish and mathematics. With the incentive program, the government expected to

> . . .give the schools an extra incentive to lift the academically disadvantaged and, e.g., develop and test new teaching methods. This will increase the focus on all children becoming as clever as possible (Danish Government, 2016, own translation).

The schools selected for participation were those with the most significant relative number of academically weak students on average over the three school years of 2013/2014, 2014/2015, and 2015/2016, conditional on at least 11 academically weak students each year. If the schools reached an improvement target compared to this 3-year average (the baseline), they would receive a bonus. The size of the bonus depended on the number of students graduating from the schools: 50 students and below yielded DKK 1.3m (EUR 175,000), between 50 and 100 students yielded DKK 1.4m (EUR 187,500), and more than 100 students yielded DKK 1.5m (EUR 200,000). The target was set to reduce the number of students performing below the threshold mark by 5%p, 10%p, and 15%p for the school years 2017/2018, 2018/2019, and 2019/2020. Although the bonus was linked to academically weak students' achievements, the definition of "weakness" was retrospectively based on exam results. Consequently, the measure solely concerned a subgroup of students, but the program implied no clear separation between academically weak students' performance and the school in general.

Participation in the program was voluntary. Out of the 127 schools that were offered participation, 104 opted in. The baselines for the share of academically weak students were 29% to 62% (an average of 40%). A few months after the exams in June (more specifically in September), the bonuses were paid. For an average school participating in the program, the bonus was about 5% to 8% of the ordinary budget.

Simultaneous with the incentive model, a supplementary project, Program for Enhancing Students (PES), was established to aid performance improvements through courses, workshops, and consultations. It was free of charge for participating schools who had to opt in to participate. The program focused on providing teachers with tools for parent-teacher collaboration, continuous evaluation and feedback, intensive learning programs, and student-to-student learning. Of the 104 schools, 87 participated in PES.

Methodology

We approached four schools participating in the FEAWS program that were selected on the basis of theoretical sampling. We selected two from the half with the highest percentage of

Table 1. Participant Schools, Characterizations, and Interviewee Positions.

School	Baseline	Interviewee positions		Results
		First year	Second year	
A	Above 40%	Principal	Principal	Won
B	Below 40%	Vice principal	Principal	Lost
C	Above 40%	Principal	Principal	Won
D	Below 40%		Vice principal	Won

academically weak students and two from the half with the lowest—one from each set having a positive and a negative value-added score (Ladd & Walsh, 2002). The Ministry of Education calculates such scores outside of the performance-funding program as the difference between a graduating year group's actual and expected marks given their socio-economic background. The sampling allowed for diversity, as the value-added score would entail that one school from each category would be under- and overperforming. However, no differences connected to the statuses were apparent between the schools in their managements' self-understandings, approaches, and results.

All four schools had between 40 and 50 graduating students, making them eligible for a bonus of DKK 1.3 m (EUR 175,000). Initially, all four schools agreed to participate in our study. One school, however, later declined, leaving us with three schools. A similar school was selected to take its place, and it agreed to participate. We conducted interviews at the original three schools twice: Once during the first year of the FEAWS program and once during the second year. We conducted a single interview at the fourth school during the second year. Table 1 summarizes the data.

Based on the literature (Ganon-Shilon & Schechter, 2019; Paletta et al., 2020; Spillane et al., 2002), we understand principals as mediators and spokespersons for school-level approaches to FEAWS. We interviewed the school managers in charge of the FEAWS project, including two principals (schools A and C) and two vice-principals (schools B and D), with one becoming an acting principal at the time of the interviews (School B). All the principals had teaching and varying management degrees, and they worked as full-time administrators without teaching obligations. Interviewing management could weaken the findings' dependability (Guba & Lincoln, 1982) on washback as an instruction-level practice, but it allows for a broader view of organizational-level decisions and consequences. Furthermore, the two sets of interviews meant that managers could get feedback from both external influences and teachers, allowing for, for instance, pushback from teachers, which would also enhance the findings' confirmability (Guba & Lincoln, 1982). To improve the credibility and authenticity of our analysis (Parker & Northcott, 2016), we shared a draft of the paper with the interviewees to allow for comments and reactions.

The empirical data comprised seven semi-structured interviews as well as documents issued by the Danish Ministry of Education describing the incentive model. The interviews were based on semi-structured interview guides (Brinkman & Kvale, 2015) to leave room for exploring the interviewees' different understandings through probing, while keeping a format that permitted comparisons. The interviews lasted approximately 90 minutes and were taped and transcribed. The first round of interviews took place between June and July 2018, whereas the second round took place about 6 months later, between January and February 2019. The interview with School D was performed in October 2018.

The interview guides had four themes. The first theme examined the principals' decision to participate in the program. The questions concerned initial perceptions of the program, for example, advantages and disadvantages to being selected and participating. We probed for potential differences between hierarchical levels at the schools and between the principal and other stakeholders, for example, parents or local municipal administrations. The second theme explored how principals experienced FEAWS, whether it affected the perceived exam-related stakes and how such perceptions differed between hierarchical levels. It contained questions about how the scheme's targets were perceived by management and teachers, probing how the scheme could affect stakes. The third theme concerned responses to the FEAWS. We asked questions about what activities had been initiated, whether the activities supported academically weak students and were meant to achieve targets, and how the principal sought to affect teachers. The fourth theme was the expected long-term implications of participation at the specific school and in general. We asked questions concerning whether the FEAWS could have unintended consequences, for example, for students not in the incentivized focus, relating the issue to expectations of management and their perceived expectations from teachers. Themes one, two, and three were mainly emphasized in the first round of interviews. At the same time, in the second round of interviews, we also asked interviewees to describe and reflect on events between interviews.

While interviewing, we were careful to avoid asking inappropriate leading questions (Brinkman & Kvale, 2015, pp. 199–201). Interviews often tend to focus on the practitioners rather than their practices. Still, that can be overcome by focusing on asking interviewees for specific activities and concrete examples as well as encouraging specificity in answers (Blossing et al., 2019), which we did. We also transcribed word for word and had the coding reviewed by all three authors while including the questions and whole interview sequences in the presentation of data to enhance the dependability of the findings and analyses (Guba & Lincoln, 1982). The central question was whether an authentic representation is provided of what has been studied. We focused on what the phenomena meant to the principals rather than their frequencies or probability,

thereby enhancing the analysis' credibility, and achieving qualitative generalizations (Parker & Northcott, 2016).

The initial analysis of the gathered data was based on two coding cycles in NVivo 12. It was an iterative analytical process (Eisenhardt, 1989) of going back and forth between the literature, coded sequences, and rereading parts of the transcripts. For the first cycle, we used concept coding for interpretations, initiatives, and washback of FEAWS along with in vivo coding for expressive statements (Saldaña, 2016). We aimed to keep the analysis grounded in the language of the interviewees to enhance the analysis' credibility (Guba & Lincoln, 1982). We developed a case summary for each school to explore general tendencies and differences between implementations based on the coding.

After the initial coding and analysis cycle, we developed a common coding scheme based on memo writing and displays (Miles et al., 2014) and the above-mentioned theoretical framework. The refined coding scheme focused on interpretations and evaluations of actions. It was meant to analytically corroborate and further explore the results of the initial analysis to increase the analysis' dependability. We based the coding scheme on concept coding and evaluation coding (Saldaña, 2016, p. 140) with codes on perceptions of the FEAWS, the FEAWS' relation to schools, the role of school management, perceived stakes, and consequences. We also concept coded washback categorized as teaching to the test, narrowing the curriculum, and focusing on marginal students based on the definition in the theoretical framework. We used the three categories as heuristics to understand the school-level effects while utilizing evaluative coding to define whether principals perceived the actions positively or negatively.

The findings are representative of the general sentiments and are expressed as the interviewees emphasized them, if not stated otherwise. We do not elaborate on initiatives specifically attributed to certain schools due to anonymity concerns. The findings section is presented such that each section is related to one of the paper's three research questions.

Findings

Principals' Perception of Performance Funding Based on Exam Results

The school managers met the FEAWS program with pragmatism, expressing sceptical but positive attitudes. Performance funding was generally not well-regarded. For example, the principal of School C found it somewhat provocative that schools were to finance initiatives, and only if they succeeded would they "get some money . . . to pay for [what] had actually been done" (C1). Others similarly emphasized that the funding principle increased financial risks by providing funding at the end rather than at the start of an initiative. The FEAWS, however, affected the schools' foci. Principal A noted that the FEAWS legitimized a focus on marks and, more specifically, academic performance previously perceived problematic because it could conflict with students' learning and personal development. Although the concerns reflected a criticism of standardized testing in literature (e.g., Amrein-Beardsley, 2009; Booher-Jennings, 2005), most interviewees perceived it as a positive change.

Even though the schools' success in attaining the bonus could influence principals' perceptions, we did not find substantial differences. Schools A, C, and D obtained the bonus after the first year. They received mostly positive reactions from stakeholders. When School A attained its bonus, its principal said, "There has been this spirit of 'yes! We succeeded'" (A1). Although characterized as a victory for the entire school, some teachers remained critical of the FEAWS. Even if School B did not reach the target, the negative consequences were limited. School B's vice-principal recounted, "Well, our educational director [in the municipality] noted it, of course. When we're in a project, we want to meet its demands. But we haven't been reprimanded or anything like that" (B2). As such, the upside of winning was greater than the downside of not achieving targets.

Nonetheless, interviewees mentioned several challenges. First, they worried that FEAWS might lead to more performance funding in the Danish educational sector. Irrespective of the perceived results, no one wanted more performance funding. As one principal reasoned, "[t]he goal is fine . . . but you could also have said that we support you with . . . [the PES], and you get the 1.3m beforehand. . . . That might have worked just as well" (A2).

Second, it was questioned whether factors not included in the model could influence results, especially student composition. Some suggested that schools with a large share of non-native Danish speaking students might not improve enough to achieve the bonus. A vice-principal argued that "[t]he challenges with that student group might be that extreme . . . that it's more like special classes" (D2).

Finally, there was the possibility of cheating. The principal of School A saw it as a theoretical possibility, but not something likely to occur:

> Some do better; some do worse on any test. You can't do anything about that. You can play with the thought of giving [DKK] 150,000 to a family to move, because they have bad children, but we don't do that in Denmark (A1).

Dishonesty was expressed to represent "a warped mentality" (A1) because the municipal schools are for all students. However, as school B's vice-principal suggested:

> If we had that one student who needed a mark of 10 at the exam [to achieve the target score]? Well, I think, don't get me wrong, but then he would get that mark. If that's correct, then there's something that's not right (B1).

Such outright cheating was considered unethical. Although vice-principal B expressed a willingness to cheat

by changing a student's mark to attain the bonus, the other interviewees would not. No specific examples of cheating were given, and interviewees only mentioned inappropriate initiatives in a hypothetical sense. Overall, the interviewees expressed that the FEAWS had driven positive changes, which contrasts with the negative portrayal of result-oriented accountability measures in literature (Hardy et al., 2019; Imsen et al., 2017; Jacobsen & Rothstein, 2015).

Perception of Stakes When Performance Funding Is Introduced

Schools differed with respect to their perceived chances of attaining the bonus. Notably, differences in student composition between year groups were believed to influence results. All four schools stated that they discerned financial risks and their chances of achieving the bonus by projecting students' expected marks. The projections affected choices relative to stakes.

All interviewees aimed at achieving the targets and bonus. However, pressure to achieve the targets differed because of the differently projected chances. School A experienced little to no pressure during the first year, expecting that the bonus would be earned easily. Principal A explained that "the class that we had in the first year was a quite skilled class, and participation was pretty much free" (A1). In turn, he summarized that

> We have probably done more testing and been a bit more focused on the individual student and some subjects than we used to . . . [but] when you ask me, have we done a lot? No (A1).

In contrast, the principal at School C decided to invest by spending more than the allocated budget on additional staff to achieve the target and in that way have the financing to create a better learning environment for the school. Investing could influence the school adversely if it failed, as that would mean dismissals. It effectively raised the stakes by creating a potential downside that, for example, was not experienced by School B when losing during the first year. The principal managed the risk by maintaining a projected "safety margin" of at least one student to gain the bonus.

During the second year, School C's economic conditions changed due to the municipality's general budget reductions, and the FEAWS-related investments were terminated. Subsequently, the principal learned that the teachers had felt excessively pressured by the incentive during the first year. They had been worried about the school's finances, something the principal had believed was solely his concern. He stated that projects had already been a contributing factor to a teacher's stress-related breakdown, and he believed that if investments had continued, more teachers would have been similarly affected. After termination of the specific initiatives, the efforts were perceived to be "more sincere" (C2) because the focus changed to "the student becoming as skilled as possible rather than [. . .] achieving the 4" (C2). The experience affected the principal's opinion about the FEAWS. During the first interview, he was very positive, but later, he became critical of the incentives and the focus on students with a GPA of approximately 4.

Schools B and D also saw its teachers being exposed to pressure. Some ninth class teachers felt increased pressure to accomplish targets, but the principals believed it was being handled and therefore not problematic. Pressure on teachers was part of the reason why the vice-principal at School B opted not to implement an investment strategy:

> I can't budget with [DKK] 1.3m that we might not get . . .. The only way to [reduce cost] is to lay off staff. And I think that would make you incredibly anxious about not reaching the targets (B2).

Accordingly, stakes were not predetermined by FEAWS' design. They were affected by other factors that could not be influenced at the school level (e.g., student group composition) and enacted by management's choices, specifically in terms of investments. Based on school C's experience and school B's reflections, investments to improve the chances of obtaining the bonus increased stakes. Even if it was solely meant to affect school management, it could affect teachers as well.

Changes and Consequences of the Performance-Based Funding Program

All schools gained insight or inspiration from the PES' courses and workshops, and they introduced initiatives based on them. Generally, the interviewees preferred PES to the FEAWS. It was elaborated that:

> The FEAWS was a catalyst to get started, I think. But, if it had just been the FEAWS on its own, it wouldn't have done it, because we wouldn't have acquired all that knowledge from the workshops . . . then we wouldn't have accomplished what we [did] . . . not at the present moment, at least (C1).

The FEAWS was a motivating force, but the PES shaped the different efforts. As mentioned by Edwards and Mindrila (2019), studies on high-performing schools have identified practices that lead to genuine improvements. These practices often result in "slow, incremental growth and require exceptional effort by educators" (Edwards & Mindrila, 2019, p. 4). Because of the FEAWS' role as a catalyst, focusing efforts, the schools had to choose the "low-hanging fruit that we most likely would be able to lift the most in six months" (B1) because the bonus required immediate improvements. For instance, serving breakfast to students was an easy way to improve students' start at the exams. It was beneficial to both the students' experience of the exams as well as their results, which were both important. As the vice-principal of School B explained, "if we reach the baseline or if we don't, the

things that we've initiated . . . [must] accomplish something" (B1). As such, the efforts were perceived beneficial, even if they did not necessarily require exceptional effort.

All four schools engaged in various forms of teaching to the test in the form of practizing test-taking and teaching test-optimizing strategies (Amrein-Beardsley, 2009; Popham, 2001). Three of the four interviewees directly mentioned teaching to the test, noting its negative connotations. For example, School A mentioned that it increased testing to assess students' expected marks and initiated intensive learning programs for students based on the tests. The principal explained that "I think that's something you always train" (A1), and FEAWS just reinforced it. All argued that their test-taking practice was legitimate because it gave the students the best chance of succeeding as part of learning how to approach tests. Holm and Kousholt (2019) also found that Danish teachers taught to the test, while admitting to its negative reputation. The principals provided specific examples of teaching to the test practices, such as when teachers at School C guided individual students' schoolwork during the year by deciding with each student in the final year classes which types of math problems they should focus on knowing how to solve.

All the principals noted giving additional attention to the students on the cusp of the GPA threshold because of the FEAWS. For instance, students were selected for intensive learning programs, and the progress of these students and their likelihood of success were discussed at management meetings. As Booher-Jennings (2005) suggested, solely focusing on marginal students can come at a cost to other student groups. The improvement target implied by the FEAWS, however, was based on the number of academically weak students among all ninth class students, not a predetermined subgroup. Consequently, singling out specific marginal students was fraught with uncertainties, which was noted and coped with by, for instance, principal C through his previously mentioned "safety margin."

Some new efforts only addressed the students that were regarded as marginal, but others included all ninth class students or all classes. For example, School B implemented specific initiatives aimed at those they considered marginal students, while generally focusing on reading, spelling, and the parts of mathematics perceived to be weak spots at the school. Initially, all Danish and mathematics teachers and the vice-principal participated in PES, but the participation was later limited to ninth class teachers due to increasing costs. It reflected a concern for efforts positively affecting all students, but the marginal students nonetheless became the most prioritized, given that their specific initiatives continued.

Most schools invested additional resources to improve the ninth class results rather than reallocate resources from other classes. School C did it by investing. School B and D took it from their budgetary slack: resources not already spent on mandated instruction that could be freely spent. Management emphasized its responsibility to handle potential dilemmas between the model's incentive and its intent to improve the entire school, thereby shouldering the responsibility for avoiding negative washback. The increased focus on marginal students at the schools, in turn, was positive because the specific students, as well as other student groups, benefited from the additional attention and new efforts.

The principals of Schools A and C and the vice-principal of School B, however, expressed that they might be focusing too much on tests. Teaching to the test could produce negative washback, although what this entailed was contested. School C's principal said that if similar tests were done "25 times, only for the sake of training, that would not be all right" (C1), whereas school B practized exactly such intensive training. School C's principal further stated that singling out students for intensive learning programs was problematic, which Schools B and D did without reservation. School D's vice-principal reasoned that such initiatives were unacceptable if they only affected short-term results yet could be permissible if not harming the student and done to achieve funding for improvements at the school.

The schools increased the relative focus on Danish and mathematics, thereby reducing the time spent on other subjects and narrowing the curriculum. They also increased their focus on subjects perceived as weak spots for that year's students. According to Au (2007) and Brill et al. (2018), such actions can have detrimental effects on the scope of learning. However, increasing the focus on Danish and mathematics was argued to be beneficial because they are fundamental subjects, implying that additional learning could positively affect other subjects as well. Furthermore, for example, at school B, their teaching to the test practices were also an extension of an existing "municipal reading initiative" (B2) rather than just because of the FEAWS.

When asked whether the efforts could have negative consequences for other student groups, the vice-principal of School B explained that:

> The easy answer would be 'yes'. What we have done is to say, these eight students . . . we'll concentrate on those. However, and this is important for me to say, that does not mean that we have compromised the learning situation for [other student groups]. But it does mean that there has been a resource allocation [prioritising those students] (B1).

This was because the resources allocated had not already been spent; that is, it was not taken away from other students. The vice-principal of School D worried that some schools could "yield amazing results in one or two years, and then no more" (D2) because they focused on "hacks you can do if you're cynical enough" (D2) to achieve targets without actually improving anything, where the interviewee noted the impreciseness of testing. The principal of School C noted that some might expect that the schools launched the initiatives only for the bonus. His choice to invest could be interpreted as such. He clearly expressed, however, that this was not his intention:

> I simply don't want anyone to say that I only did it for the money. No one should be able to say that I only helped those that were likely to get a 4. In ten years, when I look back on this project, I want to be able to look myself in the eyes (C1).

Discussion and Conclusion

Studies have shown how school management mediate policy and local practice (Lambersky, 2016; Paletta et al., 2020; Shirrell, 2016) by being positioned between potentially conflicting external and internal demands (Koyama, 2014; Shaked & Schechter, 2017; Spillane et al., 2002). This paper investigates whether performance funding based on exam results has similar consequences to other means of increasing the stakes of testing, and it demonstrates the complexities involved when principals mediate between performance funding, teachers, and daily work practices.

We find that the principals interviewed had preconceived, negative opinions about performance funding but nevertheless focused on its possibilities: They met the performance-funding program with scepticism but changed practices to reach the targets. We show that the program did not necessarily increase the exam-related stakes, although it had the potential to do so. We illustrate several initiatives that could be interpreted as teaching to the test (Popham, 2001), narrowing the curriculum (Amrein-Beardsley, 2009), and focusing on marginal students (Deming et al., 2016). Although these behaviors potentially were negative washback, we demonstrate that they were rationalized as positive by principals and often enacted as general initiatives to benefit students. In spite of the initial scepticism among the principals and the fact that much literature has been critical of result-oriented measures (Hardy et al., 2019; Imsen et al., 2017; Jacobsen & Rothstein, 2015), managers held positive attitudes toward the program's intention, content, and consequences. They mainly perceived the program as a catalyst for positive change, even if they did not want more performance funding.

Since exit exams in Denmark are considered low-stakes testing (Andersen & Nielsen, 2020), this paper focuses on whether performance funding linked to exam results increases stakes. Our findings indicate that it did. However, how and to what degree differed. The stakes and their consequences depended on the perceptions of the improvement targets. If they were perceived as easy, for example, because of the skills of the ninth class in question, the new initiatives received less managerial attention. By contrast, when the targets were perceived as challenging but achievable, pressure to improve the students' marks increased, and new activities gained additional attention. As such, the stakes depended on a school's specific situation.

The stakes also depended on the actions taken by the principals, specifically in terms of investments. In our findings, investing to achieve targets was possible by spending above the allocated budget. It would increase financial risk, thereby increasing stakes. Whereas other principals spent additional resources on the marginal students by allocating their financial slack, one chose to invest. He accepted the associated increased stakes but meant for the stakes only to affect himself and to insulate the teachers. He coped with the additional pressure through an intentional strategy of having a "safety margin" of at least one student performing above the required target, testing students to ascertain the likelihood of success. The principal, as the others, made good faith efforts (Spillane & Kenney, 2012, p. 550) to improve performance. However, the principal stated that he misinterpreted the teachers' experiences during the first year where teachers perceived the program as stressful; something not found at other schools. Rather than making the principal's life miserable, as in Spillane et al. (2002, p. 747), the teachers reacted by feeling that the efforts were insincere and did not initially make their feelings clear to the principal.

The teachers' experiences of stakes as stressful and insincere, that is, negative washback, seemed to depend on the school's investment strategy. This finding would raise the question of whether negative washback through the investment choice became qualitatively different by introducing a potential "downside" to being offered a bonus, or if it merely further increased the stakes and thus affected the teachers with a higher intensity. The fact that the principal became highly critical of the program and that other interviewees considered excessive stakes a possibility would suggest the latter. For example, another principal chose not to invest to avoid adversely affecting the teachers. This made for an initial divergence in interpretation between the two principals, which later became aligned through the investor's experiences. Both, as well as the other managers, implied that it was their responsibility to avoid negative washback.

This paper confirms the vital role of principals in interpreting and enacting new policies and accountabilities in daily practices in line with other studies (e.g., Koyama, 2014; Shaked & Schechter, 2017; Shirrell, 2016) and how strategic choices affect the understanding of current situations (Ganon-Shilon & Schechter, 2019). Moreover, we illustrate that the principals' mediating role has its limitations in terms of influencing teachers' perceptions of stakes, where the mediation does not necessarily achieve its intended effects. Coburn (2016, p. 472) questioned how policy could reshape the role of the principal. We demonstrate that performance funding based on students' exit exam results extended the role through the potential new tool of investing. We thus highlight that investments could enact stakes by the principals' own choices rather than solely owing the performance-funding scheme's design. We also illustrate that principals cannot necessarily control who will be affected by the increased stakes.

Previous research has indicated that washback effects are expected for high-stakes tests, documenting that such tests also have unintended consequences (Alderson & Wall, 1993; Hill et al., 2016; Kelley & Protsik, 1997; Lee & Medina, 2018).

Unintended consequences have also been found in performance funding in higher education (Ortagus et al., 2020; Umbricht et al., 2017). This paper contributes to the extant literature by demonstrating that similar consequences can be found within basic education when performance funding is applied. The schools' reactions to the incentives of the performance-funding model were similar to those in a high-stakes accountability system.

Contrary to the literature that often emphasizes the dysfunctionality of unintended consequences (e.g., Amrein-Beardsley, 2009; Ortagus et al., 2020; Paletta et al., 2020), this paper indicates that such behavior can be perceived as beneficial to the students. The results imply that whether teaching to the test, narrowing the curriculum, or focusing on marginal students (Amrein-Beardsley, 2009; Paletta et al., 2020) will have negative washback depends on other factors than whether specific initiatives are implemented. If washback was for the betterment of the students, interviewees considered it appropriate.

However, if initiatives were perceived to be taken to obtain bonuses, they led to insincere efforts and were deemed problematic. Teaching to the test in our case may be perceived less problematic than in the Anglo-Saxon literature because exams reflect course objectives and general skills valued by teachers (Beuchert & Nandrup, 2018) and have stakes to students as well as to educators.

Higher stakes or being at the cusp of achieving the bonus would seem to make negative washback more likely. This was especially the case in terms of cheating, which was considered a possible type of negative washback. It is well-established in the literature that the use of incentivized targets in connection with budgets can have adverse behavioral outcomes (e.g., Jensen, 2003). One interviewee suggested, albeit purely hypothetically, that he himself would cheat to achieve the bonus if presented with the opportunity. In their literature review of accountability, De Wolf and Janssens (2007) mentioned how several papers find that increased accountability based on test scores "result[s] in the exclusion of more pupils from tests" (p. 390). We did not find indications of similar unintended consequences. The reason may be that the improvement target for the FEAWS program was measured in relation to all enrolled students at the school, not only to students that attended the exit exam. Even if we cannot validate this condition, the result indicates the importance of considering a performance-funding program's specific details and accountability initiatives.

Consequently, our findings align with the literature in that high-stakes testing could cause negative washback. However, we also show that stakes are not necessarily implied by the system, as typically assumed (e.g., Au, 2007), but at least partly enacted by the school management. We contribute by showing how principals enact increased stakes of performance funding through their own choices.

In this research, we studied the organizational-level washback of performance funding based on students' exit exam results. The findings will be relevant for policymakers considering implementing performance funding within basic education in general or improving educational outcomes for specific groups of students. Our results suggest that policymakers should carefully consider the pros and cons of implementing it as a policy. Even though teaching to the test, narrowing the curriculum, and focusing on marginal students were self-perceived positive washback effects in the eyes of the principals, the possibility that the teachers experience a higher intensity pressure than the principals expected and wanted indicates how results-based performance funding can be mediated unsuccessfully, and introduce unintended consequences, such as negative washback. The willingness to cheat could also undermine the intended results. The results show the importance of communication between management and staff when dealing with funding-based measures to practitioners. Examining how performance funding affects testing outcomes and washback calls for careful consideration of the pros and cons of accountability systems, especially when high-stakes test-based funding mechanisms are introduced.

Given that this is only the second study of performance funding based on student test results after Al-Samarrai et al. (2018), more studies are needed to understand it further. The findings regarding the creation of stakes and unintended consequences and the meanings enacted around them should be studied in more depth to ascertain their effects in other ways than those mediated by the school management. We only interviewed persons with the managerial responsibility for the implementation, whereas the actors implementing the performance-funding model are the teachers. Another limitation is that the interviews were conducted with only 4 out of the 104 schools participating in FEAWS, and other schools might have experienced different results. Because Danish exit exams are valued by teachers (Beuchert & Nandrup, 2018), the transferability (Guba & Lincoln, 1982) to contexts with standardized testing is questionable. However, the findings show such a different testing context and its potential benefits in relation to washback. Furthermore, the extended role of the principal would seem to depend on performance funding rather than tests, thereby increasing the results' transferability.

Acknowledgments

An earlier version of this paper was presented at a TrygCenter seminar at Aarhus University and the EURAM 2020 Conference, The Business of now: the future starts here, 4 to 6 December 2020, Trinity College, Dublin, Ireland. The authors are grateful for the helpful comments from both audiences and from the EURAM discussant Michele Meoli. Further, we wish to thank Hans Englund, Margit Malmmose, and Allan Hansen for their helpful comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Per Nikolaj Bukh https://orcid.org/0000-0003-2374-6975

References

Akpinar, K. D., & Cakildere, B. (2013). Washback effects of high-stakes language tests of Turkey (KPDS and ÜDS) on productive and receptive skills of academic personnel, *Journal of Language and Linguistic Studies*, *9*(2), 81–94. https://doi.org/10.14198/raei.2010.23.09

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115–129. https://doi.org/10.1093/applin/14.2.115

Al-Samarrai, S., Srestha, U., Hasan, A., Nakajima, N., Santoso, S., & Wijoyo, W.H.A. (2018). Introducing a performance-based component into Jakarta's school grants: What do we know about its impact after three years? *Economics of Education Review*, *67*, 110–136. https://doi.org/10.1016/j.econedurev.2018.10.005

Amrein-Beardsley, A. (2009). The unintended, pernicious consequences of "staying the course" on the United States' no child left behind policy. *International Journal of Education Policy and Leadership*, *4*(6), 1–13. https://doi.org/10.22230/ijepl.2009v4n6a199

Andersen, S. C., & Nielsen, H. S. (2020). Learning from performance information. *Journal of Public Administration Research and Theory*, *30*(3), 415–431. https://doi.org/10.1093/jopart/muz036

Andreasen, K. E. (2019). The impact of PISA studies on education policy in a democratic perspective: The implementation of national test in Denmark. In C. Ydesen (Ed.), *The OECD's historical rise in education* (pp. 133–153). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-33799-5_7

Andreasen, K. E., Kelly, P., Kousholt, K., McNess, E., & Ydesen, C. (2015). Standardised testing in compulsory schooling in England and Denmark: A comparative study and analysis. *Bildung und Erziehung*, *68*(3), 329–348. https://doi.org/10.7788/bue-2015-0306

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*(5), 258–267. https://doi.org/10.3102/0013189x07306523

Beuchert, L. V., & Nandrup, A. B. (2018). The Danish national tests at a glance. *Danish Journal of Economics*, *156*(1), 1–37.

Biggs, J. B. (1995). Assumptions underlying new approaches to educational assessment. *Curriculum Forum*, *4*(2), 1–22.

Blankenberg, B., & Phillips, A. (2016). Performance funding in Illinois higher education: The roles of politics, budget environment, and individual actors in the process. *Educational Policy*, *30*(6), 884–915. https://doi.org/10.1177/0895904814556748

Blossing, U., Roland, P., & Sølvik, R. M. (2019). Capturing sense-made school practice. The activities of the interviewer. *Scandinavian Journal of Educational Research*, *63*(7), 1007–1021. https://doi.org/10.1080/00313831.2018.1476404

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

Brill, F. J., Grayson, K., Kuhn, L., & O'Donell, S. (2018) *What impact does accountability have on curriculum, standards and engagement in education? A literature review*. National Foundation for Educational Research.

Brinkman, S., & Kvale, S. (2015). *InterViews: An introduction to qualitative research interviewing*. Sage.

Burgess, S. M., Propper, V., Slater, H., & Wilson, D. (2005). *Who wins and who loses from school accountability? The distribution of educational gain in English secondary schools* (Discussion Paper 5248). Centre for Economic Policy Research.

Cheng, L. (1999). Changing assessment: Washback on teacher perceptions and actions. *Teaching and Teacher Education*, *15*(3), 253–271. https://doi.org/10.1016/S0742-051X(98)00046-8

Chua, C. L., & Mosha, H. J. (2015). Managing school internal mechanisms for performance improvements in secondary education: Case of six secondary schools in Eastern Zone in Tanzania. *SAGE Open*, *5*(4), 1–9. https://doi.org/10.1177%2F2158244015610172

Coburn, C. E. (2016). What's policy got to do with it? How the structure-agency debate can illuminate policy implementation. *American Journal of Education*, *122*(3), 465–475. https://doi.org/10.1086/685847

Danish Government. (2016). *Government policy* (in Danish: Regeringens politik A til Å: Folkeskolen). Copenhagen.

Danish Ministry of Education. (2017). *Funding for improving academically weak pupils in primary school* (in Danish: Pulje til løft af fagligt svage elever i folkeskolen). Retrieved July 10, 2020, from https://www.uvm.dk/puljer-udbud-og-prisuddelinger/puljer/puljeoversigt/tidligere-udmeldte-puljer/grundskole/pulje-til-loeft-af-fagligt-svage-elever-i-folkeskolen-skolepuljen

Danish Ministry of Education. (2018). *Ministerial order on funding for improving performance of academically weak pupils in primary school* (In Danish: Bekendtgørelse om pulje til løft af fagligt svage elever i folkeskolen) (BEK (117 of 21/02/2018)). Danish Ministry of Education.

Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2016). School accountability, postsecondary attainment, and earnings. *Review of Economics and Statistics*, *98*(5), 848–862. https://doi.org/10.1162/REST_a_00598

De Wolf, I. F., & Janssens, F. J. G. (2007). Effects and side effects of inspections and accountability in education: An overview of empirical studies. *Oxford Review of Education*, *33*(3), 379–396. https://doi.org/10.1080/03054980701366207

Dougherty, K. J., Jones, S. M., Lahr, H., Natow, R. S., Pheatt, L., & Reddy, V. (2016). *Performance Funding for Higher Education*. John Hopkins University.

Edwards, N. R., & Mindrila, D. L. (2019). Improving graduation rates: Legitimate practices and gaming strategies. *Education Policy Analysis Archives*, *27*(41), 1–25. https://doi.org/10.14507/epaa.27.4222

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review*, *14*(4), 532–550. https://doi.org/10.2307/258557

Englund, H., Frostenson, M., & Beime, K. S. (2019). Performative technology intensity and teacher subjectivities. *Scandinavian Journal of Educational Research*, *63*(5), 725–743. https://doi.org/10.1080/00313831.2018.1434825

Figlio, D. N., & Ladd, H. F. (2015). School accountability and student achievement. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education finance and policy* (pp. 194–210). Routledge.

Ganon-Shilon, A., & Schechter, C. (2019). School principals' sensemaking of their leadership role during reform implementation. *International Journal of Leadership in Education*, *22*(3), 279–300. https://doi.org/10.1080/13603124.2018.1450996

Gawlik, M. A. (2015). Shared sense-making: How charter school leaders ascribe meaning to accountability. *Journal of Educational Administration*, *53*(3), 393–415. https://doi.org/10.1108/JEA-08-2013-0092

Golann, J. W. (2015). The paradox of success at a no-excuses school. *Sociology of Education*, *88*(2), 103–119. https://doi.org/10.1177/0038040714567866

Guba, E. G., & Lincoln, Y. S. (1982). Epistemological and methodological bases of naturalistic inquiry. *Educational Communication and Technology*, *30*(4), 233–252.

Hardy, I., Reyes, V., & Hamid, M. O. (2019). Performative practices and 'authentic accountabilities': Targeting students, targeting learning? *International Education Journal: Comparative Perspectives*, *18*(1), 20–33.

Herbst, M. (2007). *Financing public universities: The case of performance funding*. Springer.

Hill, T., Frederiksen, B. J., Isenman, R., Rosenberg, A., & Jayaram, S. (2016). *Paying for performance: An analysis of output-based aid in education*. Results for Development Institute.

Holm, L., & Kousholt, K. B. (2019). Beyond washback effect: A multi-disciplinary approach exploring how testing becomes part of everyday school life focused on the construction of pupils' cleverness. *Annual Review of Critical Psychology*, *16*, 917–952.

Imsen, G., Blossing, U., & Moos, L. (2017). Reshaping the Nordic education model in an era of efficiency. Changes in the comprehensive school project in Denmark, Norway, and Sweden since the millennium. *Scandinavian Journal of Educational Research*, *61*(5), 568–583. https://doi.org/10.1080/00313831.2016.1172502

Jacobsen, R., & Rothstein, R. (2015). Educational goals: A public perspective. In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education finance and policy* (77–86). Routledge.'

Jennings, J. L., & Bearak, J. M. (2014). 'Teaching to the test' in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, *43*(8), 381–389. https://doi.org/10.3102/0013189X14554449

Jensen, M. (2003). Paying people to lie: The truth about the budgeting process. *European Financial Management*, *9*(3), 379–406. https://doi.org/10.1111/1468-036X.00226

Kalkan, Ü., Aksal, F. A., Gazi, Z. A., Atasoy, R., & Dağlı, G. (2020). The relationship between school administrators' leadership styles, school culture, and organizational image. *SAGE Open*, *10*(1), 1–15. https://doi.org/10.1177%2F2158244020902081

Kelley, C., & Protsik, J. (1997). Risk and reward: Perspectives on the implementation of Kentucky's school-based performance award program, *Educational Administration Quarterly*, *33*(4), 474–505. https://doi.org/10.1177/0013161X97033004004

Koyama, J. (2014). Principals as bricoleurs: Making sense and making do in an era of accountability. *Educational Administration Quarterly*, *50*(2), 279–304. https://doi.org/10.1177/0013161X13492796

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review*, *21*(1), 1–17. https://doi.org/10.1016/S0272-7757(00)00039-X

Lambersky, J. (2016). Understanding the human side of school leadership: Principals' impact on teachers' moral, self-efficacy, stress, and commitment. *Leadership and Policy in Schools*, *15*(4), 379–405. https://doi.org/10.1080/15700763.2016.1181188

Laughlin, R., Broadbent, J., Shearn, D., & Willig-Atherton, H. (1994). Absorbing LMS: The coping mechanism of a small group. *Accounting, Auditing & Accountability Journal*, *7*(1), 59–85. https://doi.org/10.1108/09513579410050407

Lee, J. D., & Medina, O. (2018). *Results-based financing in education: Learning form what works* (Results in Education for All Children (REACH)). World Bank Group.

Luxia, Q. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, *22*(2), 142–173. https://doi.org/10.1191/0265532205lt300oa

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Sage.

Myhre, H. (2021). From knowledge promotion reform to value promotion reform: Norwegian school leaders and teachers as mediators for democracy in a time of contradictions. In A. Strømmen-Baktiar & K. Timoshenko (Eds.), *Revisiting new public management and its effects: Experiences from a Norwegian context* (pp. 125–146). Waxmann Verlag.

Ortagus, J. C., Kelchen, R., Rosinger, K., & Voorhees, N. (2020). Performance-based funding in American higher education: A systematic synthesis of intended and unintended consequences. *Educational Evaluation and Policy Analysis*, *42*(4), 520–550. https://doi.org/10.3102%2F0162373720953128

Paletta, A., Ferrari, E., & Alimehmeti, G. (2020). How principals use a new accountability system to promote change in teacher practices: Evidence from Italy. *Educational Administration Quarterly*, *56*(1), 123–173. https://doi.org/10.1177/0013161X19840398

Parker, L. D., & Northcott, D. (2016). Qualitative generalising in accounting research: Concepts and strategies. *Accounting, Auditing & Accountability Journal*, *29*(6), 1100–1131. https://doi.org/10.1108/AAAJ-04-2015-2026

Popham, W. J. (2001). Teaching to the test? *Educational Leadership*, *58*(6), 16–21.

Pouncey, W. C., Ennis, L. S., Woolley, T. W., & Connell, P. H. (2013). School funding issues: State legislators and school superintendents—adversaries or allies? *SAGE Open*, *3*(2), 1–13. https://doi.org/10.1177%2F2158244013486492

Ratner, H. (2020). Europeanizing the Danish school through national testing: Standardized assessment scales and the anticipation of risky populations. *Science, Technology, and Human Values*, *45*(2), 212–234. https://doi.org/10.1177/0162243919835031

Saldaña, J. (2016). *The coding manual for qualitative researchers*. Sage.

Shaked, H., & Schechter, C. (2017). School principals as mediating agents in education reforms. *School Leadership & Management, 37*(1–2), 19–37. https://doi.org/10.1080/13632434.2016.1209182

Shirrell, M. (2016). New principals, accountability, and commitment in low-performing schools. *Journal of Educational Administration*, *54*(5), 558–574. https://doi.org/10.1108/JEA-08-2015-0069

Spillane, J. P., Diamond, J. B., Burch, P., Hallett, T., Jita, L., & Zoltners, J. (2002). Managing in the middle: School leaders and the enactment of accountability policy. *Educational Policy*, *46*(5), 731–762. https://doi.org/10.1177/089590402237311

Spillane, J. P., & Kenney, A. W. (2012). School administration in a changing education sector: The US experience. *Journal of Educational Administration*, *50*(5), 541–561. https://doi.org/10.1108/09578231211249817

Umbricht, M. R., Fernandez, F., & Ortagus, J. C. (2017). An examination of the (un)intended consequences of performance funding in higher education. *Educational Policy*, *31*(5), 643–673. https://doi.org/10.1177/0895904815614398

Urick, A., & Bowers, A. J. (2014). What are the different types of principals across the United States? A latent class analysis of principal perception of leadership. *Educational Administration Quarterly*, *50*(1), 96–134. https://doi.org/10.1177/0013161X13489019