Side effects and the enactment of accountability: results of a comparative study in two German federal states

Corrie Thiel¹

Received: 15 February 2020 / Accepted: 9 March 2021 / Published online: 4 April 2021
© The Author(s) 2021

Abstract

The paper proposes to study side effects of accountability in education within the theoretical framework of enactment research. The potential value of this approach for the study of side effects is shown by using the example of the side effect “Dependency on Expert Judgements.” Therefore, findings from the research project “Unintended Effects of Accountability in the School System” (acronym “Nefo”) are presented. The project entailed an analysis of policy documents to describe the accountability contexts under study, as well as a survey study with 2637 participating teachers and principals to examine the distribution of side effects in the no- and low-stakes contexts of the four German federal states of Berlin, Brandenburg, Thuringia, and Rhineland-Palatinate. The findings for Berlin and Thuringia are triangulated with the results from a qualitative in-depth group-discussion substudy of the Nefo project on how teachers in Berlin and Thuringia deal with accountability measures. The analysis of policy documents reveals that in Berlin, performance results gain the status of objective data that teachers need to improve their work, while teachers in Thuringia are prompted to judge by themselves the meaning of performance results for their teaching. This suggests that “Dependency on Expert Judgements” is a larger issue in Berlin than in Thuringia. However, contrary to what would be expected, the findings of the survey study indicate that the side effect plays a greater role in Thuringia than in Berlin. To explain this counterintuitive finding, teachers’ responses to standards-based accountability in Berlin and Thuringia are delineated. The paper shows that the study of meaning-making processes, which is central to the enactment framework, must not only be ignored if one tries to understand the processes that lead to the occurrence or absence of side effects. It is also important to prevent fallacies in the interpretation of survey data.

Keywords Side effects · Accountability · Enactment · Comparative study · Germany

✉ Corrie Thiel
CorrieThiel@uni-muenster.de

¹ Institute of Educational Research, University of Muenster, Georgskommende 26,
48143 Muenster, Germany

Side effects are a well-known issue in the literature on standards-based accountability in education. Many research projects that study standards-based accountability in education report findings on side effects.¹ However, in most studies, side effects are a by-product and therefore remain largely undertheorized. On the rather rare occasions when side effects are theorized, the framework of conventional implementation research² is employed. The aim of this paper is to show that studying side effects in the enactment framework instead of that of conventional implementation research could further our understanding of the occurrence, as well as the absence, of side effects. The potential of the enactment approach is explored through the example of one side effect that Johannes Bellmann, Sebastian Schweizer, and I call “Dependency on Expert Judgements” (Thiel et al., 2017).

“Dependency on Expert Judgements” covers the phenomenon that teachers in accountability contexts start to rely on results from performance measurements in their teaching, instead of on their own professional judgments. Ball (2003) hints at this when he points out that accountability measures lead to uncertainty about reasons for actions (p. 220). But if teachers are expected to teach in a manner that is highly adaptive to individual and situational circumstances — a fundamental characteristic of professional work if one follows the classic line of theorizing professionalism — they then have to rely on their own judgments (Biesta, 2015; Molander et al., 2012). In the classic line of theorizing professionalism, building on Talcott Parsons’ seminal description of medical practice as professional practice in *The Social System* (1951), professional work such as teaching is defined as “complex knowledge-based work requiring judgement in nonroutine situations” (Darling-Hammond, 1989, p. 66; similar Svensson, 2010, p. 147). In accountability contexts, a new line of theorizing has emerged that redefined professional work according to the demands of accountability systems. In this line of reasoning, to “be ‘professional’ is to have acquired a set of skills through competency-based training which enables one to deliver efficiently according to contract a customer-led service in compliance with accountability procedures collaboratively implemented and managerially assured” (Hoyle, 1995, p. 60).³ While “Dependency on Expert Judgements” appears as highly problematic in the former line of theorizing, it seems hardly problematic at all in the latter. In this paper, I take up the former line of theorizing professionalism.

To show the potential of theorizing side effects in the enactment framework, I present findings from the research project “Unintended Effects of Accountability in the School System” (acronym “Nefo”).⁴ The project was conducted from 2011 to 2014. It entailed a comprehensive analysis of policy documents related to accountability measures, as well as a survey study with 2637 participating teachers and principals, in order to examine the distribution of side effects in the no- and low-stakes

¹ For lists of side effects known in the literature, see De Wolf and Janssens 2007 and Thiel et al., 2017.

² For an analysis of the theoretical assumptions of conventional implementation research, see Spillane et al. 2002.

³ A description of this “new professionalism” is given by Julia Evetts (2011).

⁴ The research project was led by Johannes Bellmann at the University of Muenster and was funded by the German Ministry of Education and Research (BMBF).

contexts of the four German federal states of Berlin, Brandenburg, Thuringia, and Rhineland-Palatinate (Thiel et al., 2017). These results are triangulated in this paper with findings from a qualitative in-depth substudy of the Nefo project on how teachers in Berlin and Thuringia deal with accountability measures. In this substudy, group discussions with teachers in Berlin and Thuringia were conducted and analyzed (Thiel, 2019).

In Section 1 of the paper, I describe the current approaches of theorizing side effects. Section 2 provides detailed information on the methods used in the Nefo project and its substudy. In the following three sections, empirical findings from the analysis of policy documents (Section 3), the survey study (Section 4), and the qualitative substudy (Section 5) are presented. Section 3 describes the enactment of accountability in the two German federal states of Berlin and Thuringia. It will be shown that both states introduced similar accountability measures, but nevertheless, both constitute contrasting cases regarding accountability enactment. While in Berlin, performance results gain the status of objective data that teachers need to improve their work, teachers in Thuringia are themselves prompted to judge the meaning of performance results for their teaching. In view of these findings, it seems reasonable to assume that “Dependency on Expert Judgements” is a larger issue in Berlin than in Thuringia. In Section 4, I present the distribution of “Dependency on Expert Judgements” in Berlin and Thuringia. Contrary to what was expected, the findings indicate that the side effect plays a greater role in Thuringia than in Berlin. To explain this counterintuitive finding, in Section 5, I delineate teacher responses to standards-based accountability in Berlin and Thuringia. In Section 6, conclusions are drawn.

1 Theorizing side effects

On occasions where explanations for the occurrence of side effects are provided, they mostly remain within the framework of conventional implementation research. Within this framework, policy effects are conceptualized in a rather behavioristic way as reactions of educators to policy stimuli. Accordingly, researchers strive to determine which stimuli cause educators to behave in certain ways and what variables moderate this behavior (e.g., Ehren & Visscher, 2006; Jones et al., 2017). This is complemented by a conception of educators as strategic rational actors guided by self-interest (e.g., De Wolf & Janssens, 2007, p. 389 ff.). In the discussion on side effects, it is generally assumed that side effects occur when teachers either try to gain gratifications or evade sanctions.

In research on the side effects of standards-based accountability in high-stakes contexts, side effects are usually attributed to the existence of high stakes (e.g., Abrams et al., 2003; Amrein & Berliner, 2002; Baker et al., 2010; Mintrop, 2004; Ravitch, 2016). According to this view, side effects occur because teachers try to adapt their behavior to accountability demands to gain rewards or evade sanctions associated with performance results. Research in low-stakes contexts calls the claim of high-stakes as the driving factor for the occurrence of side effects into question (Thiel et al., 2017; Penninckx et al., 2016), but explanations for the occurrence of side effects usually remain within the framework of conventional

implementation research. Van Thiel and Leeuw (2002), for example, differentiate between an *unintended performance paradox* and a *deliberate performance paradox*. A performance paradox occurs when the correlation between performance indicators and actual performance is weak. In the event of an *unintended performance paradox*, actors such as teachers have to deal with minimal and elusive accountability requirements. They do so by finding strategies to gain the most benefits for themselves or the organization for which they work. In cases of the *deliberate performance paradox*, actors actively manipulate performance measurements by finding ways to make their performance look better than it actually is. In a similar vein, Bellmann, Schweizer, and Thiel proposed theorizing side effects as *adaptive behavior according to incentives*, e.g., assimilation of teaching to test demands, and as side effects, we refer to as *evasive behavior contrary to incentives*, e.g., cheating. With these distinctions, side effects are conceptualized as those of teachers' direct *reactions* to accountability demands guided by the impulse to avoid negative consequences such as a bad reputation as a teacher and/or to gain benefits such as social appraisal.

Until now, the theoretical framework of conventional implementation research has been widely uncontested in the discussion on the side effects of standards-based accountability in education. However, in the broader discussion on educational policy, the value of such frameworks for theorizing policy work in educational settings is questioned (e.g., Spillane et al., 2002; Ball et al., 2012, p. 1 ff.). It is pointed out that one must focus on contextualized sense-making processes of different actors in the education system, in order to grasp the complexities of policy work. Here, it is assumed that teachers do not merely *react* to policies in such a way as to either gain rewards or evade sanctions but *respond actively* to policies by making sense of, mediating, struggling over, ignoring, or enacting them (Ball et al., 2012, p. 3). Such responses are understood as the outcome of a situation-specific interplay of different aspects such as professional values and beliefs, the history and culture of an institution, policy biographies of key actors, and so on. Thus, it is deemed impossible to produce relatively simple linear policy-enactment models (Ball et al., 2012, p. 142).

Below, I use the terms “response” and “responding” to indicate a shift of the theoretical framework in which side effects are conceptualized. These terms are frequently used in the literature on policy enactment, but this is not unproblematic. They are well established in behavioristic psychology, e.g., in the term “stimulus–response theory.” Nevertheless, I use “response” and “responding” due to a lack of more suitable terms. However, it should be noted that both terms are understood as antonyms to “reaction” and “reacting” throughout the paper.

2 Methods

As outlined above, the paper presents the results of an analysis of policy documents, a survey, and a group discussion. In the following section, detailed information on the methods used is provided.

Table 1 Participation rates after replacement

	School participation rate	Teacher participation rate	Overall participation rate
Berlin	64.6%	39.4%	25.7%
Thuringia	83.7%	44.9%	37.6%

2.1 Analysis of policy documents

A comprehensive content analysis of a variety of policy documents, such as state legislation, publicly accessible information about accountability measures and circular letters, was conducted to describe the enactment of standards-based accountability in Berlin and Thuringia. The analysis covers the years 1999 to 2015. In 1999, Berlin started to prepare the implementation of accountability measures. In 2015, the data collection of the group-discussion study was completed. For both states, comprehensive synopses of the content analysis results were written. The description of accountability enactment in Berlin and Thuringia in Section 3 is based on these synopses.

2.2 Survey study

The main aim of the Nefo project was to study the distribution of side effects of accountability in education in the no- and low-stakes contexts of the four German federal states Berlin, Brandenburg, Thuringia, and Rhineland-Palatinate. To do so, a stratified systematic random sample of 5% of all primary schools, and 10% of all secondary schools in the four states, was drawn. All public as well as private schools were included in the sampling frame. Only schools for pupils with special needs were excluded. The sample contained four replacement schools for each originally sampled school. Within the sampled schools, the principal as well as all fully certified teachers were selected to participate in the study. Principals and teachers participated voluntarily. Thus, not all sampled schools and teachers participated in the study. In the end, 2637 questionnaires were included in the data analyses conducted in the Nefo project. Table 1 shows the participation rates in Berlin and Thuringia after including replacement schools.

The participation rates shown in Table 1 suggest that the study results are biased to a certain extent. Therefore, conclusions about the exact distribution of side effects cannot be drawn from the study data. However, the participation rates seem sufficient to explore differences in the distribution of side effects between the two federal states. Table 2 presents the total number of questionnaires from principals and teachers in Berlin and Thuringia that are included in the study.

The questionnaire contained a variety of items measuring the distribution of side effects. “Dependency on Expert Judgements” was measured with the three items: *Feedback from VERA⁵ gives me orientation in carrying out my occupation, I need VERA test results to get reliable information about the proficiency level of my students, and We need feedback from school inspections to reliably assess the quality*

⁵ “VERA” is the name of the Berlin standardized testing program. The Thuringia testing program is called “Kompetenztests.” The survey was adjusted accordingly.

Table 2 Total number of questionnaires by federal state

	Principals	Teachers	Total
Berlin	26	487	513
Thuringia	42	473	515

of work at our school. The first item asks principals and teachers if the standardized testing program VERA provides some sort of orientation for teaching. The term “orientation” is not very specific and therefore open to a variety of interpretations. Agreement with the item could indicate that the study participants view feedback from VERA as helpful *additional* information for professional decision-making, as well as that they start to align decision-making with test results *instead* of professional judgments. By contrast, the other two items are more specific. Both items ask for a “need” for performance results in order to obtain “reliable” information on students’ proficiency levels and school quality. This implies that there is a lack of such information that is not provided by other sources, such as reliable professional judgment. However, no item was employed asking in a more direct way if principals and teachers rely increasingly on expert judgments *instead* of professional judgments.

2.3 Group-discussion study

In the substudy of the Nefo project, group discussions were conducted in order to study how teachers in Berlin and Thuringia respond to accountability demands. In this study, three group discussions in Berlin and two in Thuringia were carried out. Between three and six teachers participated in the discussions. The discussions were analyzed qualitatively with the Documentary Method developed by Bohnsack (2010). Drawing on Karl Mannheim’s sociology of knowledge and Pierre Bourdieu’s theory of practice, this method aims at reconstructing the implicit habitual knowledge that structures social practices or the “modus operandi of everyday practice” (p. 101). Such implicit habitual knowledge should not be understood in an individualistic manner as the personal characteristics of individual teachers but rather as knowledge generated and acquired in social practices. Thus, the reconstructed knowledge should be understood as collectively shared. In the substudy, the group discussions were regarded as events of the social practice of teachers speaking with each other about accountability measures. In the data analysis, the focus did not lie on *what* the participants said about accountability but on *how* they negotiated the meaning of accountability measures when they spoke to each other (Thiel, 2019). In doing so, a culturalist approach was employed, i.e., an approach of explaining social action with recourse to symbolic structures of meaning (Reckwitz, 2005). Thus, the patterns of negotiating the meaning of accountability presented in Section 5 can be understood as patterns that are specific to the professional culture of teachers in Berlin and Thuringia.

Of course, the findings of this qualitative approach are not generalizable in the same way as the results of the survey. But in the methodological debate on the Documentary Method, it is assumed that findings are generalizable if implicit knowledge can be reconstructed within a variety of group discussion sequences with different topics and in different discussions. If this is the case, one can assume that the reconstructed implicit habitual knowledge structures the practice under study (Bohnsack, 2010, p. 111 f.). The patterns presented in Section 5 were reconstructed for many different sequences of the group discussions and can thus be understood as a typical way to negotiate the meaning of accountability. Due to the limited length of this paper, the main results of the group discussion analysis can only be demonstrated for one section of a group discussion with teachers in Berlin, and one of a group discussion with teachers in Thuringia.

In the original sections, the participating teachers speak German. As is common in spoken language, the word order and choice of words deviates from standard formal, written German. In the English translation, I tried to capture the way the speakers expressed themselves. This led to some phrasings and sentence structures that are uncommon in English. The punctuation marks, brackets, and the “@” are not used in a grammatically correct sense but to capture features of the spoken language such as intonation, overlapping speech, or breaks. An explanation of the signs used can be found at the end of the paper. The participants of the group discussions are labeled “D” for “debater” and numbered in order of appearance in the group discussions.

3 Enactment of accountability in Berlin and Thuringia

3.1 The ideal-type accountability model

Accountability has been internationally on the rise now for about 40 years in welfare state sectors (Hood, 1991; Evetts, 2009, p. 249 f.). Hopmann (2008) describes the rise of accountability as a transition from the *management of placements* to the *management of expectations* (p. 422 f.). At the core of the former lies the idea that the state has to provide for various ill-defined problems of its citizens such as health, education, and security, by establishing institutions in which professionals deal with these problems according to the standards of their specific profession. Thereby, good work is defined “by the professional judgement of the adequacy of what was done” (p. 424). This is gradually being superseded by the idea that setting targets, performance measurement, and holding professionals accountable for their work is the key to assuring and developing high-quality work. Good work is defined primarily by measurable performance. Hopmann emphasizes that the management of expectations is not really about output, outcome, or efficiency as some might proclaim but about meeting set expectations. Only performances that achieve this count as successes, while other performances are either considered irrelevant or deficient.

The measurement of differences between target expectations and actual performance is the driving force of management by expectations, namely standards-based

accountability. Therefore, some authors in the German debate describe standards-based accountability as a cybernetic management model (e.g., Bellmann, 2016; Herzog, 2013; Meyer-Drawe, 2009). Three elements are fundamental to cybernetic models: (1) performance targets, (2) feedback instruments that measure differences between targets and actual performance, and (3) autonomy of the managed system, so that the system can regulate itself to close gaps between targets and actual performance. To stimulate the regulation process of the system, the state can place the process under administrative control or/and create a competitive environment that pressures the systems towards good performance. However, despite all recent efforts at gaining objective, reliable, and valid knowledge about causal relationships concerning learning in the context of “evidence-based education” (Slavin, 2002), regulation processes cannot, at least not yet, be based on such data. For this reason, Bellmann (2016) suggests speaking of accountability as “data-driven management” instead of “evidence-based management.” The ideal-type accountability model is shown in Fig. 1.

This ideal-type accountability model can be enacted in very different ways, as will now be shown with the example of the two German federal states of Berlin and Thuringia.

3.2 Berlin

The starting point for the implementation of standards-based accountability in Germany at the national level is marked by the resolution *Bildungsstandards zur Sicherung der Qualität und Innovation im föderalen Wettbewerb der Länder* (Engl. “Performance Standards to Ensure Quality and Innovation in Competition between Federal States”) that the *Kultusministerkonferenz der Länder* (Engl. “Standing Conference of the Ministers of Education and Cultural Affairs in the Federal Republic of Germany”) passed on 24th May 2002. In this resolution, the introduction of performance standards for students, “Bildungsstandards,” as well as of standardized testing programs, was determined. With this resolution, the Standing Conference of the Ministers of Education and Cultural Affairs agreed on implementing two measures that are central to standards-based accountability. Performance standards comprise a set of performance targets, and standardized tests serve as feedback instruments that measure the difference between targets and actual student performances.

Fig. 1 Ideal-type accountability model

In 2003 and 2004, performance standards were formulated for grade four in the subjects German and mathematics, for grade nine in German, mathematics, and the first foreign language, and for grade ten in German, mathematics, first foreign language, biology, chemistry, and physics. Performance standards for grade twelve followed in 2012 in German, mathematics, and the first foreign language. To support the implementation of standardized testing programs by the federal states, the Institute for Quality Development in Education (IQB) was founded in 2004 to provide items for standardized testing. The federal states were obliged to implement performance standards in their curricula and to assess on a regular basis to what extent performance standards are met by the students.

Berlin started to prepare the implementation of standards-based accountability well before the Standing Conference of the Ministers of Education and Cultural Affairs passed the resolution mentioned above. From 1999 to 2004, the pilot project *Schulprogrammentwicklung und Evaluation* (Engl. “School Program Development and Evaluation”) was conducted. Schools in Berlin were invited to develop their own quality-development concept called the *Schulprogramm* (Engl.: “School Program”) and to participate in a trial run of the Berlin school inspection procedure. Schools participated voluntarily in this pilot project. After its completion, Berlin established its accountability system in three stages from 2004 to 2015.

In the first stage, covering 2004 to 2006, the prerequisites for establishing an almost ideal-type accountability system were set up. In 2004, the education act of 1980 was replaced by a new education act signaling a paradigm shift in school governance. §9 of the new education act authorized the Ministry for Education, Youth and Family to pass statutory regulation regarding quality assurance and evaluation of schools, without the discussion and approval of the Berlin parliament. §9 has also obliged schools to continue quality development and to execute internal as well as external evaluations. Schools have been legally obliged to participate in external evaluations such as standardized testing programs and school inspections.

In addition to the performance standards formulated by the Standing Conference of the Ministers of Education and Cultural Affairs, Berlin has been setting target expectations in a quality framework comprising 16 school quality criteria. For each of the 16 criteria, quality indicators were defined in the framework.

Furthermore, the publicly accessible school database *Schulporträt* was set up in 2005 to publish quantitative as well as qualitative data about each school, including performance results. The official aim of that database was to provide each user with an objective view of individual schools. Such a database is an important prerequisite for fostering competition between schools.

In 2006, regulation *VergleichsVO* (Engl. “Comparison Regulation”) was passed by the Ministry for Education, Youth and Family and specified the implementation of standardized tests in Berlin. This marks the beginning of the second stage of accountability in Berlin, covering the years 2006 to 2011. The school authority was entrusted with administering standardized tests. To administer them, the Institute for School Quality of the States Berlin and Brandenburg (ISQ) was founded in 2006. It was assigned the tasks of analyzing the tests and providing reference values for the school authority and schools, in order to judge the test results for individual schools. In the school year 2007/2008, the standardized testing program “VERA 3” was introduced, measuring the proficiency level of students in grade 3, and in

2008/2009, “VERA 8” was established, measuring the proficiency level of students in grade 8. According to the *VergleichsVO*, detailed test results only had to be conveyed to pupils and parents. In addition, the *Schulkonferenz* (Engl. “School Conference”) had to be informed about aggregated test results of different learning groups within a school. Up to 2011, the schools had only been allowed to publish their tests results if the *Schulkonferenz* decided to do so with a two-third majority.

This development has been paralleled by the first round of mandatory school inspection for all schools in Berlin. The school inspection procedure that was introduced is highly standardized. Each inspection team consists of a school administrator, a school principal, a teacher, and a volunteer without a background in school administration or teaching. The inspection team studies school documentation, the results of an online survey provided by the ISQ, asking teachers, pupils, parents, and other pedagogical professionals working in the school about their views of the school, interviews different actors in the school, and visits at least 70% of teachers during lessons for about 20 min each. Based on the collected data, the quality criteria outlined in the quality framework defined in the first stage of the implementation of the Berlin accountability system are evaluated on a scale ranging from “A” (highly developed) to “D” (poorly developed). In addition, each quality criterion is rated on a scale ranging from “a” (the school’s average is at least one standard deviation above the average of Berlin schools) to “d” (the school’s average is at least one standard deviation below the average of Berlin schools). The results of the school inspection are reported to the school principal, the *Schuladministration* (Engl. “School Administration”), as well as the municipal council of the school. If a school performs poorly in the inspection, the school is labeled *Schule mit erheblichem Entwicklungsbedarf* (Engl. “School in Considerable Need of Development”). In this case, the period between school inspections is shortened from 5 to 2 years.

In the second stage of the implementation of accountability in Berlin, the school program has become a central instrument for data-driven administrative control of schools. In 2008, an administrative directive was passed, in which schools are obliged to update their school program with reference to the results of standardized tests and school inspections. The adjusted school program must be approved by the *Schuladministration*. In contrast to data-driven administrative control, data-driven competition between schools had not yet actively been promoted in Berlin. This changed in 2011.

In 2011, the *VergleichsVO* was replaced by the regulation *SchulQualSiEvalVO* heralding the third stage of the establishment of accountability in Berlin. The *SchulQualSiEvalVO* contains all the regulations of the *VergleichsVO*, but regulations concerning the publication of test and inspection results have been added. The school authority has been authorized to publish aggregated performance results. To do so, the database *Schulporträt*, which was established in the first stage, is used. Accordingly, data-driven competition between schools in Berlin has been made possible, especially between secondary schools. For primary schools, catchment areas have been assigned and parents have had to file an application if they have wanted their child to visit a school other than the one assigned to their child. For secondary schools, parents have been enabled to freely choose the school for their child.

Figure 2 shows the accountability system established in Berlin between 1999 and 2015.

Fig. 2 The accountability system enacted in Berlin

In the Berlin accountability system, performance results gain the status of indisputable objective measurement in a twofold sense. Once the results become part of data-driven administrative control and/or parents start to choose the school for their child based on performance results, these results become an inevitable social reality for both schools and teachers. They become “objects” in the social practice of schooling. In this social practice, performance results are treated as if they reveal the truth about teaching. This status of performance results is also reflected in two comprehensive brochures about the standardized testing programs “VERA 3” and “VERA 8,” published by the ISQ targeting teachers. Here, the use of the tests is outlined as follows⁶:

- (1) View from outside: Standardized tests *provide teachers with profound knowledge about the skills of pupils*. VERA test items cover a wide range of different requirements and thus *give teachers the opportunity to compare their own judgement about the students with the help of an objective scientific instrument*. (...)
- (2) Reflection on the effectiveness of one’s own teaching: The analysis of test results *provides teachers with feedback about the success of their teaching*. (...)
- (3) Strengthening one’s own diagnostic competences: The frequencies of solving standardized test items shows that students occasionally solve tasks by themselves which they in fact did not — or just rudimentarily — discuss in school. Therefore, test results can be an additional component of diagnosing competencies.

In all three paragraphs, test results are regarded as revealing important information to teachers concerning their teaching. The tests are thereby explicitly framed as “objective scientific instruments.” This conveys the message that teachers need such

⁶ The following is a translation of parts of page 12 of the brochures about the standardized testing programs “VERA 3” and “VERA 8.” The brochures are largely identical. They can be retrieved from https://www.isq-bb.de/wordpress/wp-content/uploads/2016/05/isq_vera3_2012-03-29.pdf and https://www.isq-bb.de/wordpress/wp-content/uploads/2016/05/Broschuere_BLN_VERA_8_2011.pdf. Emphasis added by me.

information and are not able to judge the skills of their pupils and the success of their teaching validly without them.

3.3 Thuringia

While Berlin has been seeking to enhance the quality of education by providing objective data that illuminate the reality in schools, Thuringia has primarily been trying to develop the quality of education by qualifying the teachers themselves to initiate local quality-development projects in their schools. To do so, Thuringia has implemented various development programs such as the *Entwicklungsprogramm für Unterricht und Lernqualität E.U.L.E.* (Engl.: “Development Program for Teaching and Learning Quality”), which was conducted from 2004 to 2014. Its aim was to improve the quality of education by consolidating and expanding teacher abilities to instruct in a way that promotes understanding. A core idea was to set up professional learning communities to mobilize teacher experience for further local quality-development endeavors. This idea is central to the development projects implemented in Thuringia. Accountability measures have been implemented in Thuringia as part of this broader quality-development strategy. Two stages of accountability implementation can be identified.

During the first stage, covering the years 2001 to 2008, all accountability measures during the period analyzed were introduced. As early as the school year 2001/2002, standardized tests for pupils in classes three and six were implemented. These tests became mandatory for all schools in the school year 2002/2003. Later, standardized tests for pupils in grade eight were introduced. In 2002, a project group was founded at the University of Jena that was assigned the tasks of administering and analyzing the tests. The test results were reported back to the participating teachers and schools. The teachers teaching one subject, as well as the *Schulkonferenz*, were summoned to discuss the test results, write down the results of the discussions, and stipulate development measures. Up to 2015, the test results were not reported to actors outside school.

In addition, the development project *Eigenverantwortliche Schule und schulische Evaluation EVAS* (Engl.: “Independent School and School Evaluation”) was conducted from 2005 to 2009. As part of this development project, Thuringia undertook a trial run of a school inspection procedure with schools participating voluntarily in the program. The steps undertaken in the school inspection are in general similar to the school inspection procedure in Berlin. However, there is a crucial difference regarding the assessment of the schools by the school inspection team. As in Berlin, a school quality framework was developed in Thuringia in 2006. But in contrast to the one in Berlin, this framework covers only 1.5 pages and differentiates roughly between “context quality,” “process quality,” and “effect quality.” No quality indicators are named. This framework thus does not serve as a basis for a highly standardized rating of school quality. Rather, it is only suitable for informing such a rating. Accordingly, schools in Thuringia have been receiving only written feedback.

Up to 2008, quality development was left to the schools and not administratively controlled. This changed with the modification of the Thuringia school act in 2008, which marks the start of the second stage of accountability implementation in Thuringia.

In contrast to Berlin, Thuringia did not pass a new school act but modified the existing one. In 2008, §40b, with the title *Selbstständige Schule und Evaluation* (Engl.: “Independent School and School Evaluation”), became part of the school act. This paragraph is equivalent to §9 of the Berlin school act. With this paragraph, schools have been obliged by law to continue quality development and assurance. §40b (3) states that schools must participate in inspections and conclude target agreements with the school authority. With the instrument of target agreements, school quality development has been placed under administrative control.

Up to 2015, schools have been able to decide themselves whether or not to publish their results in standardized tests and the report of the school inspection on their website. In contrast to Berlin, Thuringia has not been fostering data-driven competition between schools to enhance the quality of education. Figure 3 shows the accountability system established in Thuringia between 2001 and 2015.

As Fig. 3 shows, Thuringia introduced accountability measures similar to those in Berlin between 2001 and 2015. Thus, the difference between the two federal states does not lie so much in the implemented measures, but in the way standards-based accountability is enacted. Accountability measures are embedded in Thuringia’s strategy for developing the quality of education by mobilizing teaching experience for further local development efforts of teachers. This changes the meaning of accountability measures compared to Berlin. In accordance with this broader quality-development strategy, teachers are requested to discuss the results of standardized tests and school inspections and to agree on development measures. In Thuringia, feedback seems to be one more way of encouraging teachers to initiate their own quality-development measures. The difference between Berlin and Thuringia is striking if one compares the framing of standardized test results in the brochures about VERA 3 and VERA 8 in Berlin, with the framing of results in Thuringia. When teachers in Thuringia receive their test results, they also receive supporting information from the University of Jena on how to interpret test results. Teachers are addressed as experts who are the only ones able to assess what variables of the tested class or school played a role regarding the test outcome and to what extent

Fig. 3 The accountability system enacted in Thuringia

quality-development measures can be deduced. It is also emphasized in the information that if a school's performance is below average, this can — but does not need to — mean that the quality of teaching at this school is lower than in other schools. While in Berlin, performance results are regarded as objective data that disclose different aspects of teaching to teachers, the same results are framed in Thuringia as scientific data in need of interpretation by teachers.

Against the backdrop of the analysis of policy documents, it is reasonable to assume that the side effect “Dependency on Expert Judgements” plays a larger role in Berlin than in Thuringia. To test this assumption, results of the survey study conducted in the Nefo project will be presented in the next section.

4 “Dependency on Expert Judgements” in Berlin and Thuringia

Table 3 shows the response distribution for the three items measuring “Dependency on Expert Judgements” in the Nefo project.

As shown, the ratings from teachers and principals on items measuring “Dependency on Expert Judgements” differed greatly between Berlin and Thuringia. A statistically significant higher number of teachers and principals in Thuringia than in Berlin agreed either completely or partly with all three items: 53.7% of teachers and principals in Thuringia, but only 17.3% in Berlin, agreed with the item *Feedback from VERA gives me orientation in carrying out my occupation*; 34.5% of teachers and principals in Thuringia, but only 9.6% in Berlin, agreed with the item *I need VERA-test results to get reliable information about the proficiency level of my students* and 44.7% of teachers and principals in Thuringia, but only 24.2% in Berlin, agreed with the item *We need feedback from school inspections to reliably assess the quality of work at our school*. It is also interesting that the majority of principals and teachers in Berlin not only disagreed but did so completely with the items *Feedback from VERA gives me orientation in carrying out my occupation* and *I need VERA-test results to get reliable information about the proficiency level of my students*, while a relatively small number of principals and teachers in Thuringia did so. Teachers in Berlin seem to have a stronger aversion to standardized tests than teachers in Thuringia.

The results of the survey study suggest that the almost ideal-type accountability system in Berlin has not been fostering “Dependency on Expert Judgements” to the same degree as the accountability system that actively promotes teacher judgment. To understand this counterintuitive finding, findings on how teachers in Berlin and Thuringia respond to messages conveyed by accountability measures are presented in the next section.

5 Teachers’ responses to accountability in Berlin and Thuringia

5.1 Berlin

The discussions conducted in Berlin were characterized in general by a collective problematization of standards-based accountability. This is expressed metaphorically by one teacher in the statement: “a common enemy unites people.” But the

Table 3 Response rates for items measuring “Dependency on Expert Judgements”

Item	State	Responses				N
		Agree completely	Agree partly	Disagree partly	Disagree completely	
<i>Feedback from VERA gives me orientation in carrying out my occupation.</i>	BE*	1.8 (0.01)	15.5 (0.02)	34.5 (2.8)	48.2 (0.03)	233
	TH	10.7 (0.03)	43.0 (0.03)	34.1 (0.02)	12.3 (0.02)	348
<i>I need VERA test results to get reliable information about the proficiency level of my students.</i>	BE*	2.6 (0.01)	7.2 (0.02)	29.7 (0.03)	60.5 (0.04)	234
	TH	5.4 (0.02)	29.1 (0.03)	42.1 (0.03)	23.5 (0.03)	348
<i>We need feedback from school inspections to reliably assess the quality of work at our school.</i>	BE*	3.9 (0.01)	20.3 (0.02)	47.0 (0.03)	28.8 (0.03)	454
	TH	5.3 (0.01)	39.4 (0.03)	44.6 (0.04)	10.8 (0.02)	457

Note. BE, Berlin; TH, Thuringia; information in percent; () standard errors in parenthesis; a weighted distribution is presented; *statistically significant differences compared to Thuringia ($p \leq 0.05$); for significance testing, the response categories “agree completely” and “agree partly” were combined into one category “agreement” and the response categories “disagree partly” and “disagree completely” were combined into one category “disagreement”.

more crucial finding concerning understanding the occurrence of the side effect “Dependency on Expert Judgements” is that discussing accountability with each other becomes a social practice of collectively claiming teaching to be the area of expertise and sphere of competence of teachers. The following is illustrative of this:

D1: Ehm (1) the question is (,) diagnosis. so when I (,) normally have a class, I actually always diagnose. (...) so for me one lesson is actually enough. or one double lesson. and I can roughly the pupils (,) or I can tell the pupils where their problems are. and from this ehm also follows what I do. and whether I do that now (,) somehow (,) in lovely 32-page booklets ehh and type everything and give feedback to the parents and type it in the internet, I get the same result (,) considerably faster. (...) so (,) for me a double lesson is enough in English. in which I go through eh different things, and see how the students speak, how they read, how they ehm (,) in case of need write, this might be the next lesson, and then (,) you see certain things where the problems lie (...)

D2: But {all at once could give them maybe plainly tasks,}

D1: {Yes and eh yes ehm mmh}

D2: And see how they get solved. yes, and (,) then you see ehm (,) did they do this before, did they understand it or who didn't, or where

D1: Mhm,

D2: Yes. where are the (,) deficits, (1) this is a subject- (,) individually manageable, right,

D3: I don't believe that you can do that in a double lesson, { @ }

D2: { @ } {well}

D1: {Well} {if yes then it is ve- very crude}

D3: {With thirty pupils and the many} competencies you have then, I believe you do need a bit longer but I believe it is really about the point, the the (,) variation of teaching and (,) ehm how you see pupils (,) ehm results then may they be oral or written (,) you get clues on whatever subject. where the weaknesses of the individuals are (,) (...) I also believe, and this is the main point of critique teachers have that they always say (,) ehm “people assume, with the introduction, of this VERA eight, that I am not able to diagnose by myself.” I hear that over and over again, and it is a point there. Because in the end I do know my pupils exactly I know exactly who is not able to do mental arithmetic who is not able to draw properly who is not able to solve equations or whatever. this (,) this you simply know.

In the section cited above, three teachers talk to each other about the standardized testing program VERA. At the beginning of the section, the topic “diagnosis” is set. This is introduced as a “question”. Here, “diagnosis” is specified as a topic open to debate. Speaker D1 takes the position of someone who is not only very experienced in diagnosing (“I actually always diagnose”) but also very efficient (“for me one lesson is enough. or one double lesson”) and successful (“I can tell the pupils where their problems are”). Diagnosing is presented here as an integral part of the speaker's teaching practice. This is confirmed with the statement that diagnosing is the foundation of the speaker's decisions about what and how to teach (“from this

ehm also follows what I do”). As integral part of teaching, diagnosing is neither a new task nor one of teaching with selective importance.

Subsequently, the issue of instruments for diagnosing learning problems that are external to teaching, such as VERA, is broached by speaker D1 when he speaks about “32-page booklets”. He compares this way of diagnosing learning problems with his own method and claims the latter to be more efficient (“considerably faster”) while — according to D1 — both ways lead to the same results. Pointing out that both ways lead to the same results serves, in the conversation, as proof of the validity and reliability of the speaker’s method of diagnosing.

At the end of his turn, the speaker claims again the ability to diagnose learning problems. This claim is generalized to teachers as a whole in the expression “you see certain things where the problems lie”. At this point, the verb “see” is used twice to describe the diagnostic process. “Seeing” denotes both a direct sensual perception of a situation and grasping the meaning of something. Compared to “seeing where problems lie,” diagnosing learning problems with standardized testing material is much more indirect.

In the adjacent turn, speaker D2 amplifies what has been said by D1 by proposing to give tasks to pupils to “see how they get solved”. He again uses the verb “see” twice to refer to diagnosing. He also continues speaking about diagnosing in a generalized manner, as is the case at the end of speaker D1’s turn. One could say that D1 and D2 speak with one voice about diagnosing.

In the last part of the section cited above, D3 joins the conversation. Her turn starts with an objection to what was said by D1. She doubts that one double lesson is enough to diagnose all the learning problems of a class. But as D3 proceeds, it becomes clear that the objection should be understood as a restraint on what was said about the time needed for diagnosing, but not a general objection to the position taken by D1 and D2. On the contrary, D3 elaborates the position of D1 and D2, by repeating that teachers in general are able to diagnose learning problems in their day-to-day practice.

In the last part of speaker D3’s turn, it becomes clear that the position taken by D1, D2, and D3 can be understood as a response to the accountability system enacted in Berlin. D3 positions herself (“I also believe”) towards the standardized testing program by quoting a critique made by other teachers. In doing so, it becomes clear that D3 takes the same position as the quoted teachers. The “main point of critique” which teachers express is that the implementation of the standardized testing program VERA 8 implies that teachers are themselves incapable of diagnosing. With “and it is a point there”, D3 emphasizes that the cited perception of teachers is valid and thereby confirms that view as one which is shared by teachers in general. She explains (“because”) the critique of teachers in the following section by claiming precise knowledge about the learning deficits of her students.

In the section presented, the three teachers talking to each other take a collective stand against VERA, by claiming that they are well able to diagnose learning problems by themselves. Teachers in Berlin respond to the message conveyed by the brochures on VERA that they do need objective scientific data to become aware of the skills and learning deficits of their pupils, with the collective self-assurance that they, as well as teachers in general, gain such knowledge in their day-to-day

practice. One could also say that the accountability system enacted in Berlin provokes a strong response from teachers claiming the value of their own expertise in teaching, including judging the learning problems of their students.

5.2 Thuringia

In contrast to the discussions conducted in Berlin, those carried out with teachers in Thuringia were characterized by opposing positions regarding standards-based accountability. While some teachers pointed out that accountability measures are very helpful, others found them very problematic. Regarding the matter discussed here, it is crucial how teachers negotiated the meaning of accountability in this situation; they discussed to what extent accountability measures can contribute to their specific working situation. They positioned themselves regarding accountability measures by judging their value for their own situation. This is illustrated by the following group discussion section:

D3: I probably wrote eight competency tests by now (...), had them from the beginning (...) it developed great (...) I personally think it is a great instrument. (...) a good opportunity, it is more work of course, you sit for a while (...) but for the class I think it's good, also when you get the results for parental work, it is a great thing. I for example rated my stuff prepared everything and then my parents came, got an appointment, and went through it with their child intensely (...) and then "och we coul- we should look at this again and at this and that" this was (,) I (,) actually do that {every time like this}

D5?: {Okay. mmh mhm.}

D3: Think it is good and I know now (,) where I can continue working when I get my results.

(...)

D5: {The question is of course too and that is the regional difference. so in X}

Dx: {Yes, yes. yes. (,) mmh.}

D5: At the parents' evening eh (2) I do not reach a lot of parents. and {precisely}

D?: {Hmh}

D5: My specialists there they are certainly not there. that is that difference. right, so that is the school capture area (,) or the social environment of the school plays of course a very big role there. and whether I tell by parents what happens there, they do not care about it.

In the first turn, speaker D3 introduces herself as someone with substantial experience with standardized tests, called "competency tests" in Thuringia. In doing so, she adopts the status of an expert in conducting tests. Provided with the authority of an expert, D3 rates the tests extremely positively as "a great instrument", "a good opportunity", "good", and "a great thing". Below, she balances the efforts caused by the tests with its gains. She admits that teachers have more work with the tests, but this is outweighed by the gains regarding collaboration with parents and her own teaching. To illustrate the former, speaker D3 presents

an episode of her teaching practice. This is in line with her self-introduction as a teacher with substantial experience in standardized tests. The episodic character of the section is highlighted by her use of direct speech to present parents' reactions to tests. According to the direct speech, the reaction of parents consisted of the insight that they should go through certain topics with their child again. In the episode presented, the parents become teaching assistants by repeating the topics not yet understood by their child at home. Parents and teachers act in concert regarding learning by students. With the expression "I (,) actually do that {every time like this}", it is made clear that the episode described is by no means a unique and in fact a fairly typical case. This is commented on by another speaker — presumably D5 — signaling active listening with "Okay. mmh mhm".

Speaker D3 finishes her turn by emphasizing again that in her opinion, tests are "good". This is followed by a sentence that stands in great contrast to the expressions the teachers in Berlin used in discussing standardized tests: "and I know now (,) where I can continue working when I get my results". In addition to the gains regarding collaboration with parents, speaker D3 highlights the value of the tests for herself as a teacher. The tests are described by speaker D3 as a solid foundation for deducing what steps need to be taken next.

Below, D5 comments on what has been said by D3. It is important to note that she does not call into question what was said by D3, but she constrains its applicability by pointing at "regional differences". This is immediately supported by another speaker with "Yes, yes. yes". Similar to D3, D5 talks about her collaboration with parents. In contrast to the collaboration described by D3, this one is characterized by problems encountered in contacting parents ("I do not reach a lot of parents") and parental indifference regarding issues such as standardized tests ("they do not care about it"). This is attributed in D5's turn to the social environment of the school.

With speaker D5's turn, the meaning of standardized tests for teaching is disclosed as context-specific. Under certain circumstances — a school located in an area where parents are interested in issues concerning the school of their children — standardized tests might have the positive effects pointed out by D3. But this is specific to certain schools. Regarding other schools, the described positive effects of standardized tests cannot be found.

In the section cited above, speakers D3 and D5 discuss the potential value of standardized tests in the light of their own working situation. While the test might prove valuable in the specific situation of one teacher, it might prove irrelevant or even problematic in another. When discussing accountability measures, teachers in Thuringia constantly judge the value of the instruments for their specific situation. In the end, the measures are neither considered to be good nor problematic in general by teachers in Thuringia.

6 Conclusion

Teachers in Berlin are confronted with an accountability system that requires them to rely on evaluation results, instead of their own judgments, in order to promote the quality of their teaching. Evaluation results are framed as objective scientific data

that can illuminate various aspects of teaching. As the data from the survey study show, this is not — as one might assume — associated with a comparatively wide distribution of the side effect “Dependency on Expert Judgements”. This counterintuitive finding can be understood by analyzing how teachers in Berlin collectively respond to accountability demands when they talk to each other. They respond to the message that they need objective scientific data, with the collective self-assurance that they gain all the information needed in their day-to-day work. In countering the perception that they are not able to diagnose learning deficits of their students, they argue that teaching is their area of expertise and their sphere of competence, so that they know what is necessary.

Teachers in Thuringia on the other hand face a broader quality-development strategy in which teacher judgment — including about performance results — is actively fostered. The data from the survey study suggest that this form of accountability is associated with a comparatively wide distribution of “Dependency on Expert Judgements”. This seems to be paradoxical at first sight. But if one considers the differences between the enactment of standards-based accountability in Berlin and Thuringia, it becomes clear that accountability *means* something different in the two states. While in Berlin, the meaning of performance results is defined by an *opposition* of objective scientific data and teacher judgment, performance results are framed in Thuringia as scientific data *in need* of teacher judgment. Thus, when teachers in Berlin agree with the item *Feedback from VERA gives me orientation in carrying out my occupation*, they agree with an item asking whether they are in need of expert judgments in opposition to professional judgments. When teachers in Thuringia agree with the item, they agree with an item asking whether test results whose relevance they judge by themselves can give them orientation in their practice. Such differences relating to the meaning of survey items can only be discovered if different forms of accountability enactment are studied. This indicates that survey items actually acquire their meaning through the context in which they are administered. One and the same item can be understood very differently in different contexts, due to the experiences the actors have in answering the item. If such differences are not taken into account, this can lead to an inadequate interpretation of survey study results. The analysis of the group discussions shows that it is not the case that teachers in Thuringia collectively rely on performance results as the survey data suggest. They respond to accountability demands by judging whether or not data are helpful in their specific situation and thereby make their own judgments. Thus, the results of the survey study need to be called into question. If principals and teachers really understand the items employed to measure “Dependency in Expert Judgements” in the way just described, the items do not measure the side effect as intended in Thuringia. If this is the case, “Dependency on Expert Judgements” might play a similarly minor role in Thuringia as in Berlin.

I have presented findings concerning the side effect “Dependency on Expert Judgements”, in order to draw attention to the potential for explanations of the emergence of side effects to extend beyond the rather behavioristic explanation that such effects occur because teachers either unintentionally or deliberately try to gain rewards or evade sanctions associated with performance results. Taking teachers’

responses to accountability into account can significantly promote our understanding of the occurrence, as well as the absence, of accountability side effects in education. To demonstrate the potential of such an approach, I triangulated the findings of an analysis of policy documents, a survey, and a group discussion. In doing so, I regarded and treated the findings as being linked to each other. Although I think it is reasonable to do so, the link between the findings is only assumed and not scientifically proven. In addition, only a small number of discussions were conducted in the group-discussion study. It is possible that a higher number of discussions would have led to a more differentiated picture. Furthermore, the data were collected between 2013 and 2015. During these years, the implementation of accountability was still in progress in the two states under study. Thus, it is not clear whether or not the findings presented are specific to the implementation period of accountability in education. Despite these limitations of the triangulative approach used in the paper, I wish to emphasize three points that should be born in mind in further studies on side effects.

First, to prevent a one-sided approach to the side effects of accountability in education, one should be careful not only to study side effects within the policy framework underlying accountability. The ideal-type accountability model presented in Section 3 presupposes that teachers *react* to performance feedback by attempting to close the gaps between targets and actual performance. The possibility that accountability measures gain a different meaning in different contexts, as well as the possibility that teachers do not necessarily react but respond to the messages conveyed by accountability measures in different ways, including positioning themselves collectively towards them, is excluded from the outset. If such a positioning process emerges, it is regarded as problematic misbehavior of teachers against the backdrop of the ideal-type accountability model. The findings of the group discussions presented in Section 5 show that teachers do negotiate the meaning of accountability when talking to each other and take a stand on it.

Second, according to the findings presented above, the *emergence as well as the absence* of side effects is mediated by complex social processes that should be taken into account, so as to gain a deep understanding of side effects. Even if an accountability system is established that has all the features associated with side effects, this does not mean that such an accountability system does produce the expected side effects. As the case of Berlin shows, such accountability systems can even provoke social processes that lead to counter effects.

Third, it might be a losing battle to search for features of accountability systems that enhance or reduce side effects *in general*. In this paper, I presented findings on the occurrence of “Dependency on Expert Judgements” in the two low-stakes contexts of Berlin and Thuringia. However, these findings cannot be generalized to other side effects. The survey study conducted in the Nefo project shows that other side effects, such as different forms of cheating, are widely distributed in Berlin and in Thuringia (Bellmann et al., 2016; Thiel et al., 2017). It could be the case that teachers in Berlin and in Thuringia try to gain social prestige or fear blaming and shaming. But it could also be the case that teachers cheat for pedagogical reasons. Questions like this are open for further research.

Transcription rules {Also}: Overlapping speech; (,): Short stop; :: Slightly falling intonation; .: Strongly falling intonation; (1): Pause with a minimum length of one second; length in seconds in brackets; .: Slightly rising intonation; @: Laughter; (...): Omission

Funding Open Access funding enabled and organized by Projekt DEAL. The research presented in the paper was funded by the German Ministry of Education and Research (BMBF).

Data availability The survey data as well as the full group-discussion transcripts are available in the database “Verbund Forschungsdaten Bildung.”

Declarations

Conflict of interest The author declares no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into Practice*, 42(1), 18–29.
- Amrein, A. L., & Berliner, D. C. (2002). *An analysis of some unintended and negative consequences of high-stakes testing*. Arizona State University.
- Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., et al. (2010). *Problems with the use of student test scores to evaluate teachers*. Economic Policy Institute.
- Ball, S. J. (2003). The teacher's soul and the terrors of performativity. *Journal of Education Policy*, 18(2), 215–228.
- Ball, S. J., Maguire, M., & Braun, A. (2012). *How schools do policy. Policy enactments in secondary schools*. Routledge.
- Bellmann, J. (2016). Datengetrieben und/oder evidenzbasiert? Wirkungsmechanismen bildungspolitischer Steuerungsansätze. *Zeitschrift für Erziehungswissenschaft*, 19(Sonderheft 31), 147–161.
- Bellmann, J., Schweizer, S., & Thiel, C. (2016). Nebenfolgen Neuer Steuerung unter Bedingungen von "low-stakes" und "no-stakes". Qualitative und quantitative Befunde einer Untersuchung in vier Bundesländern. In BMBF (Ed.), *Steuerung im Bildungssystem. Implementation und Wirkung neuer Steuerungsinstrumente im Schulwesen* (pp. 208–238). Berlin: M&E Druckhaus.
- Biesta, G. (2015). How does a competent teacher become a good teacher? On judgement, wisdom and virtuosity in teaching and teacher education. In R. Heilbronn & L. Foreman-Peck (Eds.), *Philosophical perspectives on the future of teacher education* (pp. 3–22). Wiley Blackwell.
- Bohnsack, R. (2010). Documentary method and group discussions. In R. Bohnsack, N. Pfaff, & W. Weller (Eds.), *Qualitative analysis and documentary method in international educational research* (pp. 99–124). Barbara Budrich.
- Darling-Hammond, L. (1989). Accountability for professional practice. *Teachers College Record*, 91(1), 59–80.

- 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.
- Ehren, M. C. M., & Visscher, A. J. (2006). Towards a theory on the impact of school inspections. *British Journal of Educational Studies*, 54(1), 51–72.
- Evetts, J. (2009). New professionalism and new public management: Changes, continuities and consequences. *Comparative Sociology*, 8(2), 247–266.
- Evetts, J. (2011). A new professionalism? Challenges and opportunities. *Current Sociology*, 59(4), 406–422.
- Herzog, W. (2013). *Bildungsstandards. Eine kritische Einführung*. Kohlhammer.
- Hood, C. (1991). A public management for all seasons? *Public Administration*, 69, 3–19.
- Hopmann, S. T. (2008). No child, no school, no state left behind: Schooling in the age of accountability. *Journal of Curriculum Studies*, 40(4), 417–456.
- Hoyle, E. (1995). Changing conceptions of a profession. In H. Busher & R. Saran (Eds.), *Managing teachers as professionals in schools* (pp. 59–70). Kogan Page.
- Jones, K. L., Tymms, P., Kemethofer, D., O'Hara, J., McNamara, G., Huber, S., et al. (2017). The unintended consequences of school inspection: The prevalence of inspection side-effects in Austria, the Czech Republic, England, Ireland, the Netherlands, Sweden, and Switzerland. *Oxford Review of Education*, 43(6), 805–822.
- Meyer-Drawe, K. (2009). “Sich einschalten”. Anmerkungen zum Prozess der Selbststeuerung. In U. Lange, S. Rahn, W. Seitter, & R. Körzel (Eds.), *Steuerungsprobleme im Bildungswesen. Festschrift für Klaus Harney* (pp. 19–34). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Mintrop, H. (2004). *Schools on probation: How accountability works (and doesn't work)*. Teachers College Press.
- Molander, A., Grimen, H., & Eriksen, E. O. (2012). Professional discretion and accountability in the welfare state. *Journal of Applied Philosophy*, 29(3), 214–230.
- Parsons, T. (1951). *The social system*. Collier Macmillan Limited & The Free Press.
- Penninckx, M., Vanhoof, J., De Maeyer, S., & Van Petegem, P. (2016). Enquiry into the side effects of school inspection in a “low-stakes” inspection context. *Research Papers in Education*, 31(4), 462–482.
- Ravitch, D. (2016). *The death and life of the great American school system: How testing and choice are undermining education*. Basic Books.
- Reckwitz, A. (2005). Toward a theory of social practices: A development in culturalist theorizing. In G. M. Spiegel (Ed.), *Practicing history: New directions in historical writing after the linguistic turn* (pp. 245–263). Routledge.
- Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, 37(7), 15–21.
- Spillane, J., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and re-focusing implementation research. *Review of Educational Research*, 72(3), 387–431.
- Svensson, L. G. (2010). Professions, organizations, collegiality and accountability. In L. G. Svensson & J. Evetts (Eds.), *Sociology of professions* (pp. 145–166). Daidalos AB.
- Thiel, C. (2019). *Lehrerhandeln zwischen Neuer Steuerung und Fallarbeit. Professionstheoretische und empirische Analysen zu einem umstrittenen Verhältnis*. Wiesbaden: Springer VS.
- Thiel, C., Bellmann, J., & Schweizer, S. (2017). Rethinking side effects of accountability in education. Insights from a Multiple Methods Study in four German School Systems. *Education Policy Analysis Archives*, 25(93), 1–27.
- van Thiel, S., & Leeuw, F. L. (2002). The performance paradox in the public sector. *Public Performance & Management Review*, 25(3), 267–281.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.