Education Policy Analysis
Archives/Archivos Analíticos de Políticas
Educativas
ISSN: 1068-2341
EPAA@asu.edu
Arizona State University
Estados Unidos

Thiel, Corrie; Schweizer, Sebastian; Bellmann, Johannes

Rethinking Side Effects of Accountability in Education: Insights from a Multiple Methods Study in Four German School Systems

Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, vol. 25, 2017, pp. 1-29
Arizona State University
Arizona, Estados Unidos

SPECIAL ISSUE
Global Perspectives on High-Stakes Teacher Accountability Policies

education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 25 Number 93 August 21, 2017 ISSN 1068-2341

Rethinking Side Effects of Accountability in Education: Insights from a Multiple Methods Study in Four German School Systems

Corrie Thiel

Sebastian Schweizer
&
Johannes Bellmann
University of Muenster
Germany

Citation: Thiel, C., Schweizer, S., & Bellmann, J. (2017). Rethinking side effects of accountability in education: Insights from a multiple methods study in four German school systems. *Education Policy Analysis Archives, 25*(93). http://dx.doi.org/10.14507/epaa.25.2662 This article is part of the special issue, *Global Perspectives on High-Stakes Teacher Accountability Policies*, guest edited by Jessica Holloway, Tore Bernt Sorensen, and Antoni Verger.

Abstract: Based on a research project comprising data from an interview study and a survey with teachers and school principals in four German federal states (*Bundesländer*), this paper questions the claim that the side effects of accountability in education are bound to high-stakes contexts, and also provides evidence of side effects occurring in no- and low-stakes contexts. The findings suggest that side effects cannot be fully explained by certain implementation features of accountability regimes (e.g. high stakes), but should rather be understood as a result of implementation features as well as

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 9/8/2016
Revisions received: 7/8/2017
Accepted: 7/8/2017

systematic effects of accountability in education.
Keywords: policy analysis; accountability; high stakes tests; comparative study

Repensando los efectos secundarios de la rendición de cuentas en la educación. Resultados de un estudio múlti-métodos en cuatro sistemas de la escuela alemana
Resumen: Basado en un proyecto de investigación que comprende datos de entrevistas y encuestas con profesores y directores de escuelas en cuatro Estados federados (Bundesländer), este trabajo cuestiona la afirmación de que los efectos secundarios de rendición de cuentas en educación están limitados a contextos de gran repercusión (high-stakes) y también proporciona evidencias de efectos secundarios que ocurren en contextos de baja o sin repercussion (low-stakes). Los resultados sugieren que los efectos secundarios no pueden explicarse completamente por ciertas características de la aplicación de los regímenes de rendición de cuentas por ejemplo de gran repercusión, pero debe entenderse más bien como ambos: como resultado de características de implementación y como efectos sistemáticos de rendición de cuentas en la educación.
Palabras-clave: Análisis de la política; rendición de cuentas; pruebas de gran repercusión; estudio comparativo

Repensando efeitos colaterais de responsabilização/accountability na educação. Resultados de um estudo multi-métodos em quatro sistemas escolares alemães
Resumo: Baseado em um projeto de pesquisa que abrange dados de entrevistas e questionários com professores e diretores de escolas em quatro unidades da federação alemã (Bundesländer), este texto questiona a alegação de que os efeitos colaterais de responsabilização/accountability na educação estão vinculados aos contextos onde há sanções (high-stakes) e também traz evidências dos efeitos que ocorrem em contextos com menores ou sem consequencias (low-stakes). Os resultados sugerem que efeitos colaterais não podem ser totalmente explicados por certas características de implementação de regimes de responsabilidade, por exemplo, high-stakes, mas antes devem ser entendidos como ambos: como resultado de características de implementação e como efeitos sistemáticos de responsabilização na educação.
Palavras-chave: análise política, responsabilidade/accountability; high-stakes testing; estudo comparativo

Introduction

Side effects are frequently observed phenomena in the Anglo-American discussion on accountability in education. In surveying the literature, one can identify an entire set of "unintended consequences" (Jones, Jones, & Hargrove, 2003) or "collateral damage" (Nichols & Berliner, 2007) of such systems. The phenomena described range from different forms of behavior such as the infamous "teaching to the test", "cheating" or "cream skimming" to a variety of side effects referring to teacher attitudes, such as "erosion of trust" or "deprofessionalization". Based on literature reviews, Bellmann and Weiß (2009) and de Wolf and Janssens (2007) both present lists of such side effects.

In the Anglo-American discussion, side effects are usually associated with the existence of high stakes (Abrams, Pedulla, & Madaus, 2003; Amrein & Berliner, 2002; Mintrop, 2004). One of the best-known examples of high-stakes accountability is the No Child Left Behind Act (No Child Left Behind [NCLB], 2002), whose system of incentives and sanctions is blamed for numerous side effects (Baker et al., 2010; Ravitch, 2016). This has recently been confirmed by the Committee on Incentives and Test-Based Accountability in Public Education, which has reviewed various test-based incentive programs (Hout & Elliott, 2011). Some research projects have even attempted to

establish a direct link between the level of high-stakes pressure and the dissemination of side effects in certain federal states (Nichols, Glass, & Berliner, 2005; Pedulla et al. 2003).

In the aftermath of PISA 2000, the German federal states (*Bundesländer*) established accountability measures in their education systems, following developments in the United States and Great Britain. However, because the experiences with high-stakes accountability in these countries repeatedly served as a 'cautionary tale' (Böttcher, 2012), accountability measures were introduced in the German federal states without linking high stakes to those measures. Thus, accountability regimes in German federal school systems can be characterized as no- and low-stakes contexts. In high-stakes contexts, consequences in the form of rewards or sanctions are linked to evaluation results. In no-stakes contexts, evaluation results serve as descriptive information only (Nichols & Berliner 2007; Ryan & Feller 2009). Concerns that the implementation of tests for the purpose of feedback and comparison, without positive or negative consequences, may be associated with side effects are countered in Germany with the remark that side effects are only a problem of high stakes (Blömeke, 2004). In short, until now, side effects have not been considered a noteworthy problem, as long as high stakes are not introduced. As a result, the question of whether the side effects of accountability systems in education can also be observed under conditions of no-stakes or low-stakes has received almost no attention (Jäger, 2012).

Despite all variations, accountability systems are essentially about the combination of instruments of standards-based accountability (educational standards, standards-based testing, school inspections) and choice policies (school autonomy, extended options in school choice, enhanced competition through private schools) (e.g. Betebenner, Howe, & Foster, 2005; Wells & Holme, 2005). This combination is expected to produce significant impulses for quality development in the education system. Both groups of instruments provide feedback for schools, informing them about their performance level (in comparison to other schools), either bureaucratically through standards-based tests and inspections or through market signals. Such feedback can serve as an information basis for teachers and principals in decision-making and may be understood as an incentive to enhance performance simultaneously. For many participants in educational policy, this combination of standards-based accountability and choice policies ranks as a blueprint for successful school systems. According to a White Paper of the Department for Education in the UK (DfE, 2010), the most innovative and high-performing school systems are those which combine high levels of autonomy for both teachers and schools with high levels of accountability.

The implementation of standards-based accountability and choice policies in Germany's federal states without the introduction of high stakes provides a good opportunity to question the claim that side effects are a matter of high stakes only. The research project 'Unintended Effects of Accountability in the School System' (acronym Nefo), funded by the German Ministry of Education and Research (BMBF), examined side effects in no- and low-stakes contexts of four German federal states. An interview study with 101 teachers and principals and a survey with 2637 respondents in Berlin, Brandenburg, Rhineland-Palatinate and Thuringia were conducted. Contrary to the expectation that side effects do not pose a substantial problem in no- and low-stake contexts, the findings of the Nefo project indicate that side effects are a serious issue. The findings suggest that the side effects of accountability are not only a matter of high stakes and therefore a matter of implementation characteristics of the accountability systems introduced, but should be understood as systematic effects. To clarify and validate these assumptions, the paper proceeds as follows:

In the first part, we outline the theoretical concept of side effects. We then present our approach of studying side effects. Subsequently, we introduce a list of side effects that includes both side effects known from the Anglo-American discussion of accountability in education and supplementary phenomena indicated by teachers in the interview study of the Nefo project. We also

comment on some side effects in the list. In the following section, we describe some results of our survey study. These results suggest that a distinction between adaptive behavior (according to incentives) and evasive behavior (contrary to incentives) is advantageous for studying side effects of accountability in education. The paper concludes with propositions of how to explain the occurrence of side effects of accountability in education and with suggestions for further research.

Theoretical Concept

To discuss the problematic effects of accountability in education, a wide variety of concepts is introduced in the literature, e.g. side effects, unintended consequences, undesirable outcomes, external effects, unintentional effects and "effets pervers" (van Thiel & Leeuw, 2003, p. 271), each with different implications for how phenomena are classified as side effects.

To clarify the understanding of side effects in this paper, the distinction between intention and function is crucial. As McLaughlin (2002) points out, functional approaches prescind from intentionality and focus on functions to explain social phenomena. Concepts such as unintended consequences, undesirable outcomes and unintentional effects suggest the existence of definable intentions. The study of side effects with reference to intentions rests on two assumptions: (1) events such as the introduction of accountability systems in education can be traced back to an identifiable person or group of persons and (2) it is possible to identify these individuals' or group's intentions validly and reliably. Thus, before observable phenomena can be classified as unintended consequences, undesirable outcomes or unintentional effects, the intentions associated with establishing accountability need to be identified. Such an attempt encounters severe problems. Usually, multiple actors play a part in policy making, thus different intentions might be associated with one policy. Tackling this issue, one faces the problem of "re-intentionalizing" certain effects, meaning that effects are viewed as intended only in retrospect.

Concepts such as external effects, "effets pervers" and side effects elude such problems. Instead of identifying side effects in relation to intentions, the (ascribed) function of an activity or of an institutional arrangement can be used to classify observable phenomena. Following official statements from policy makers, the function of accountability in education can be summarized as enhancing the quality of education, which is usually specified in terms of efficiency and equity (Woessmann & Schuetz, 2006). In a nutshell, accountability in education is expected to directly and efficiently enhance the results of educational systems while simultaneously advancing educational equality. Therefore, side effects can be defined as those effects that are in some way dysfunctional for enhancing the quality of education (Ehren & Visscher, 2006). The task for researchers studying side effects is thus to demonstrate how and why an effect might adversely affect efficiency and equity.

While the above-mentioned considerations apply to the research on side effects in general, previous approaches to side effects allow conclusions for a theory of side effects specific for accountability in education. As early as 1964, Etzioni outlined certain problems when it comes to measuring the effectiveness and efficiency of organizations:

> Most organizations under pressure to be rational are eager to measure their efficiency. Curiously, the very effort – the desire to establish how we are doing and to find ways of improving if we are not doing as well as we ought to – often has quite undesired effects from the point of view of the organizational goals. Frequent measuring can distort the organizational efforts because, as a rule, some aspects of its output are more measurable than others. Frequent measuring tends to encourage over-production of highly measurable items and neglect of less measurable ones. (p. 9)

Especially in the case of organizations with vague goals, such as schools, "the distortion consequences of over-measuring are larger when it is impossible or impractical to quantify the more central, substantive output of an organization" (p. 9). According to Etzioni (1964), three insights can be gained for a theory of accountability side effects in education: First, side effects do not necessarily emerge from a "wrong" way of dealing with accountability instruments by actors such as teachers; instead, accountability systems might rather be systematically associated with the risk of producing side effect. Second, side effects come into play not only when actors try to avoid pressure that is associated with performance measurement (as it is the case with different forms of "cheating", "window dressing" and "cream skimming"), but also when they learn how to adapt their behavior to the measured demands. In doing so, non-measured demands face a high risk of losing importance in everyday practice (as it is the case for different forms of "reallocation", "coaching" and" myopia") (van Thiel & Leeuw, 2002). Third, it is necessary to take into account that the introduction of performance indicators in organizations characterized by so-called ill-defined problems bares a high risk. This is especially true for professional behavioral contexts that are hallmarked by both an orientation towards vaguely defined societal target values and a high degree of uncertainty, as is the case with practice in schools (Frey & Osterloh, 2006; Hopmann, 2003; Levin, 1998).

Another substantial contribution to a theory of side effects of performance indicators and feedback systems stems from Campbell (1975), who describes side effects as phenomena of corruption. He describes the social mechanism underlying side effects as follows:

> The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor. (p. 35)

Corruption can be differentiated at this point; on the one hand, corruption can be used with reference to indicators. The indicators operationalizing quality can lose their validity as soon as actors narrow their performances to the defined indicators. Koretz (2005) points to this when he speaks of score inflation in the course of standards-based accountability (Hout & Elliott, 2011). On the other hand, one can identify the corruption of actors. The introduction of performance indicators can lead to a restructuring of an individual motivation in the sense that the intrinsic motivation based on professional beliefs is crowded out in favor of external rewards or commands (Frey, 2012). This is especially true if the individuals affected perceive the intervention as controlling (Deci, 1971; Frey, 2012).

Following Campbells' theorizing of side effects as corruption, Nichols and Berliner (2005) describe the side effects of accountability systems in education as "corruption of indicators and educators". But contrary to Campbell, they regard the occurrence of high stakes linked to performance results as the fundamental cause of the production of side effects. Unfortunately, this supports the assumption that high stakes are the ultimate reason for the appearance of side effects, instead of taking into account that side effects might be associated systematically with performance measurement as a means of quality assessment. Literature and research from Germany, as well as research conducted in the Anglo-American context, commonly narrow the discussion on side effects to the issue of high stakes, which leaves limited room for both alternative explanations and the development of a general theory of side effects. As early as 1956, Ridgway suggested not only concentrating on indicators that entail serious consequences for all stakeholders, but also to consider the "dysfunctional consequences of performance measurement" (p. 240), even when performance measurement is used only to obtain feedback and information.

In order to elaborate on the distinction between side effects as a corruption of indicators and of educators, the approach of van Thiel and Leeuw (2002) can be adopted. Referring to a broad sample of empirical studies, they observe a "performance paradox" (p. 271) when it comes to measuring performance. For them, the term "performance paradox" addresses a weak correlation between performance indicators and actual performance. Even though van Thiel and Leeuw focus only on side effects as the corruption of indicators in their approach, they do point out that side effects can come into play either unintentionally or deliberately. Thus, van Thiel and Leeuw distinguish between an "unintended performance paradox" (p. 272) and a "deliberate performance paradox" (p. 274). While the unintended performance paradox is basically related to insufficient or imprecise accountability requirements, the deliberate performance paradox represents forms of behavior that are very close to sabotage in an attempt to raise performance. In their view, the deliberate performance paradox presupposes the unintended performance paradox.

Van Thiel and Leeuw (2002) consider the side effects of accountability in education to be a problem of shortcomings in performance measurement. For them, the occurrence of side effects is therefore caused by minimal accountability requirements, the elusiveness of policy objectives, goals that are unquantifiable and thus hard to measure, and a strong emphasis on monitoring and efficiency within organizations. The crucial problem of accountability in public sector provision is seen as the absence of appropriate and valid performance indicators "that minimize the dysfunctional effects and maximize functional effects" (p. 277). According to the intentional paradigm, they do not consider the possibility that problematic effects might be systematically linked to accountability systems. As a result, a rather normative perspective is applied that entails the danger of focusing solely on avoiding a performance paradox by improving the monitoring system. However, the distinction between the unintended and deliberate performance paradox of van Thiel and Leeuw stresses that side effects do not necessarily occur as a result of conscious misbehavior. Side effects might also emerge unnoticed, caused by features of accountability other than high stakes.

In summary, three conclusions can be drawn with regard to side effects: First, in the debate over accountability in education, side effects are attributed to the existence of high stakes. Although the degree to which side effects occur might depend on specific implementation features of the accountability systems introduced, e.g. high stakes, the existence of side effects may be understood more as systematic effects of accountability in education. Second, it seems obvious that the corruption of indicators, e.g. "teaching to the test" that results in score inflation, is a substantial problem if one strives to raise quality via standards-based testing. Yet, it is less obvious why the effects of accountability measures that do not distort indicators, but have consequences for educators and educational practice, prove to be problematic for enhancing educational quality as well. Taking the corruption of educators into account broadens the view on the problematic effects of accountability in education. Third, side effects might not only occur when teachers try to avoid accountability pressures. They might also emerge when teachers adapt to measurement demands. This emphasizes the point that some side effects emerge because teachers conform to the measurement system, and other side effects arise because teachers oppose measurement requirements. These considerations suggest the need for a broadened view of accountability side effects in education, which takes into account that side effects may not only be a problem that occurs when the implementation features of accountability regimes lead to a corruption of indicators that teachers employ when trying to avoid pressure. A broadened view has consequences for the study of side effects. In the following section, we outline the approach to side effects as applied in the Nefo study.

Methods

For the reasons outlined above, we chose to rely on the concept of side effects rather than "unintended consequences", although the project title suggests otherwise. As a working definition, we understand the side effects of accountability in education as impacts that are negative or ambivalent with respect to efficiency and equity. That is to say, we are not trying to measure educational quality itself, but rather to identify potential constraints to a complex process of quality development. Whether or not the identified effects do in fact interfere with quality development is open to debate. Following this approach, no formal definition of side effects can be employed. Instead, to classify effects of accountability measures as side effects, a sound argument has to be provided as to how these effects might impact on efficiency and equity. Hence, there is neither a definitive list of side effects of accountability in education, nor are phenomena arbitrarily characterized as side effects.

To demonstrate this approach of classifying phenomena as side effects, we refer briefly to a phenomenon discussed in the literature as "cream skimming" (Whitty & Power, 2000). In order to ensure high test scores in standards-based tests, schools try to recruit high-performing pupils, while simultaneously attempting to get rid of pupils with low performance records. From the perspective of a school, this can be regarded as a rational behavior. But it is questionable whether this behavior really raises the efficiency of a school system, as one cannot be sure that the good test results of high-performing schools compensate for those of low-performing schools. Such behavior can also be a threat to equity, because social segregation of students might be promoted (Levin, 1998; Moe, 1995).

Employing this approach to side effects, two research questions were examined in the Nefo study: What kind of side effects can be found in no- and low-stakes contexts and how far are the side effects spread in the no- and low-stakes contexts in the four federal states under study? To answer these questions, three steps were taken, consisting of (1) an analysis of a variety of policy documents in order to describe the accountability regimes of Berlin, Brandenburg, Rhineland-Palatinate and Thuringia, (2) an interview study and (3) a survey study. In the following, we briefly sketch out these steps.

The Analysis of Policy Documents

The four federal states Berlin, Brandenburg, Rhineland-Palatinate and Thuringia were chosen to investigate side effects because prior research indicated that these states differ in the way accountability measures were introduced (Döbert, Rürup & Dedering 2008; Rürup 2007). However, to get a detailed picture of their accountability measures, a comprehensive analysis of policy documents, such as state legislation, publicly accessible information about accountability measures, circular letters etc., was conducted. Based on the information collected throughout the documentary analysis, a detailed description of the accountability regime in each federal state was prepared. This was the basis for developing the footprints for each accountability regime presented in the next section.

It is characteristic of accountability regimes to link bureaucratic and competition-based regulation of schools and professional self-control by means of data, such as results of performance tests and of school inspections (Bellmann, Schweizer & Thiel, 2016; Thiel, Cortina & Pant, 2014). Hence, the focus of the analysis of policy documents laid on the question of whether and how evaluation data were integrated into bureaucratic and market-based regulation and professional self-control in the federal states. For the four states, the emphasis placed on each of the three modes of data-driven regulation was rated to be "low", "medium" or "high".

The emphasis on market-based regulation was rated "high" if a relatively high degree of school autonomy could be found in a state, school choice was introduced, and evaluation data were made publicly available via a database. If only two of the three measures could be found, the emphasis on data-driven market-based regulation was rated "medium" and if only one or none of the three measures was found it was rated "low". There was one exception. If school autonomy is evident, as well as school choice, whereas evaluation data are not made publicly available at the system level, one cannot speak of data-driven competition. If this was the case, we rated the strength of market-based regulation as "low".

The strength of data-driven bureaucratic regulation was rated by the number of evaluation instruments for whose very procedure it is inherent that the evaluation results are received not only by schools, but also discussed with administrators. If three such evaluation instruments could be found, the strength was rated "high", if two instruments were found it was rated "medium", and if only one or no instrument was found, it was rated "low". Examples for evaluation instruments introduced in the four states are performance tests, school inspections and internal evaluation instruments.

In a similar way, the emphasis on professional self-control was rated by the number of evaluation instruments that made teachers deal with the results themselves, without being embedded in market-based or bureaucratic modes of regulation.

This approach seemed to be a reasonable way to capture the most crucial differences at the "policy level" (Tyack & Cuban, 1995, p. 40) of the four federal states, but there are limitations to such an approach. It leads to a rather crude description of accountability regimes, neglecting more subtle differences, which might turn out to be crucial in the implementation of data-based regulation, e.g. ignoring different degrees of obligations to meet target agreements between administrations and schools. In addition, the descriptions do not refer to "policy implementation" (p. 40). It might be the case, for instance, that the emphasis on market-based regulation was rated "high", but in reality, there was no effective competition, because the schools are too far apart from each other to be actual competitors.

The comparative design of the study provides a background for the interpretation of results of the quantitative study presented below. Accordingly, differences between federal states are traced back to differences in the accountability regimes of the four states, although causal relations cannot be tested statistically in a cross-sectional study like Nefo. Thus, the explorative approach led to the generation of hypotheses concerning factors relevant for the occurrence of accountability side effects in education that need to be tested in further research.

The Interview Study

An interview study was carried out in the four states under study. Taking the list of side effects published by Bellmann and Weiß in 2009 as a starting point, we conducted the interviews to explore whether or not the side effects known from the Anglo-American debate on high-stakes accountability can also be found in the no- and low-stake contexts of Berlin, Brandenburg, Rhineland-Palatinate and Thuringia. In addition, the interview study was undertaken to detect further side effects and add them to the list.

To reduce bias as much as possible, principals and teachers of primary as well as all major kinds of secondary schools in the four states were interviewed. In the four states under study, primary schools are comprehensive. After leaving primary school, pupils attend usually one out of three kinds of secondary schools. They can visit a comprehensive school, a secondary school that leads to a grade 10 school leaving certificate qualifying for vocational training, or a Gymnasium that leads to a university entrance diploma. We aimed at interviewing the principal and two teachers, one

of them actively engaging in school development processes and the other one being uninvolved, of three schools per school type in each state. A private school was added to the sample for comparison reasons. Unfortunately, it was not possible to reach the intended number of interviews in all cases. Table 1 gives an overview over the total number of interviews carried out in the Nefo study. The schools were recruited via telephone calls with the school principals. All schools participated voluntarily in the study.

Table 1
Number of Interviews by Federal State

	BE	BB	RP	TH
Primary School	5	9	9	10
Secondary School up to grade 10	3	9	9	10
Gymnasium	3	9	13	9
Comprehensive School	3	0	0	0
Total	14	27	31	29

Note. BE = Berlin; BB = Brandenburg; RP = Rhineland-Palatinate; TH=Thuringia.

A total of 101 semi-structured interviews were carried out. For this purpose, an interview guide was developed by the project team consisting of rather broad questions concerning how accountability measures are perceived and dealt with, and of rather narrow questions concerning specific side effects known from the Anglo-American debate.[1] The interviews had a length between 23 and 122 minutes. All the interviews were audio taped.

In order to answer the research questions central for the interview study, the interviews were analyzed by two members of the project team, listening to the audio tapes of the interviews. Their task was to write down every occasion in which an interviewee drew a causal relation between accountability measures and a phenomenon occurring in everyday school life. Following this step, it was discussed by the project team whether or not the identified phenomena could be classified as side effects, using the approach outlined at the beginning of this section. As a result, some side effects were added to the list presented by Bellmann and Weiß in 2009.

The Survey Study

The survey study was based on the extended list of side effects. It was conducted to examine the distribution of side effects in the no- and low-stakes contexts of the four states. The IEA Data Processing Center (DPC) in Hamburg provided a stratified systematic random sample of 5% of all primary schools and 10% of all secondary schools in the four federal states. The sample included public as well as private schools. Only special schools for handicapped people were excluded from the sampling frame. The DPC provided a sample containing four replacement schools for each sampled school. Within the sampled schools, the principals as well as the total of fully certified

[1] An English version of the interview guide used in the Nefo study is available at http://www.uni-muenster.de/EW/en/forschung/projekte/nefo/materialien.html.

teachers were supposed to participate in the study. But as schools as well as principals and teachers participated voluntarily, it was not possible to ensure that the criteria set in the sampling plan were met in all cases. Table 2 presents the participation rates at the school level, at the teacher level (including principals and teachers) and the overall participation rate. All participation rates refer to participation after including replacement schools.

Table 2
Participation Rates After Replacement

	School Participation Rate	Teacher Participation Rate	Overall Participation Rate
BE	64,6%	39,4%	25,7%
BB	83,3%	50,3%	41,9%
RP	75,9%	55,8%	42,3%
TH	83,7%	44,9%	37,6%

Note. BE = Berlin; BB = Brandenburg; RP = Rhineland-Palatinate; TH=Thuringia.

The participation rates shown in Table 2 seem sufficient to explore the question of whether or not side effects pose a substantial problem in no- and low-stakes contexts. However, they do suggest that the results are biased to a certain extent, so that conclusions about the precise distribution of side effects have to be drawn with caution. The final number of questionnaires included in the study is presented in Table 3. A total number of 2637 participants responded to the survey.

Table 3
Total Number of Questionnaires by Federal State

	BE	BB	RP	TH
Principals	26	40	77	42
Teachers	487	538	954	473
Total	513	578	1031	515

Note. BE = Berlin; BB = Brandenburg; RP = Rhineland-Palatinate; TH=Thuringia.

Two questionnaires were applied in the Nefo study. The core questionnaire, consisting of items asking for side effects, was administered to both principals and teachers.[2] In addition, a complementary questionnaire, containing items asking about school characteristics, was given to the school principals. Both questionnaires were developed by the project team. It was intended to cover the various side effects presented in Table 4.

[2] An English version of the questionnaire used in the Nefo study is available at http://www.uni-muenster.de/EW/en/forschung/projekte/nefo/materialien.html.

Such an approach is of course at the expense of studying individual side effects in depth, but seemed suitable for testing the prevailing assumption that side effects do not pose a substantial problem in contexts where accountability measures are not linked to high stakes.

The Accountability Regimes of the Four States under Study

The documentary analysis has shown that Berlin, Brandenburg, Rhineland-Palatinate and Thuringia have shared a shift to accountability measures during the past 15 years. In all states performance tests and school inspections were introduced. However, the four states do form a contrasting sample concerning the emphasis and interplay of bureaucratic and competition-based modes of regulation as well as professional self-control.

At the level of policy action, the accountability system of Berlin is characterized by a high emphasis on market-based and bureaucratic regulation. In Berlin, output data are made available through a public data base, schools have relatively high levels of autonomy and parents are free to choose their preferred secondary school. Secondary schools are entitled to decide about the enrollment of pupils in the case of excess demand. The results of external evaluations like school inspections and performance tests as well as the results of internal evaluation procedures are subject to target agreements between school administration and schools. In contrast to market-based and bureaucratic regulation, little emphasis is placed on professional self-control. Only the results of pupils evaluating their teachers serve exclusively as feedback for teachers. These evaluations of teachers by pupils are neither part of the bureaucratic control of schools nor part of market-based regulation. Figure 1 visualizes the strength of professional self-control, bureaucratic regulation and market-based regulation in Berlin.

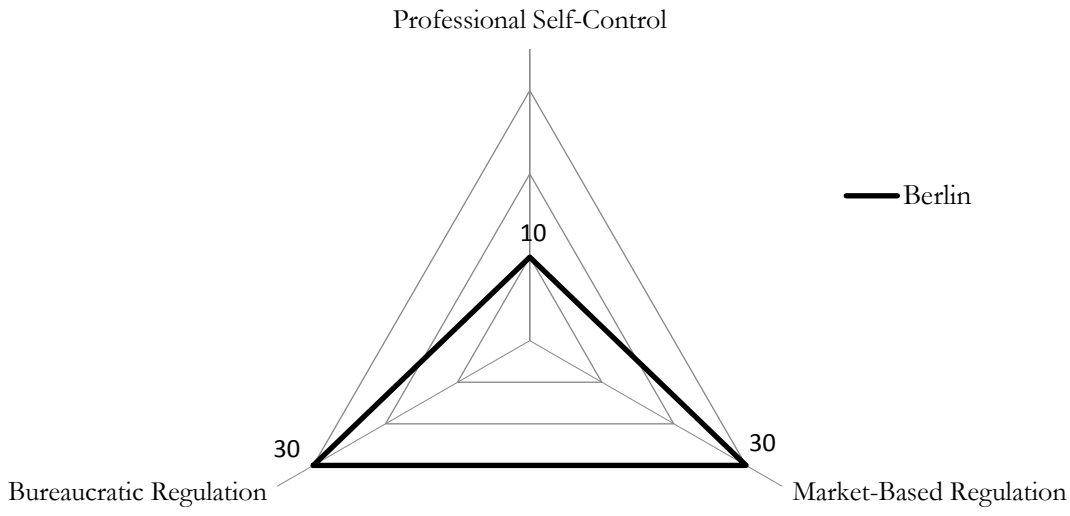

Figure 1. Strength of professional self-control, of bureaucratic regulation and of market-based regulation in the accountability regime of Berlin. "10" indicates that the strength of the regulation mode was rated "low", "20" that it was rated "medium" and "30" that it was rated "high".

Brandenburg has introduced the same accountability measures as Berlin during the last 15 years. Differences between Berlin and Brandenburg occur less at the level of policy action, but might be fundamental at the level of policy implementation. It can be assumed that market and competitive pressures play a more important role in Berlin, because it is a city state and Brandenburg a territorial state. Figure 2 shows the strength of the different modes of regulation in Brandenburg.

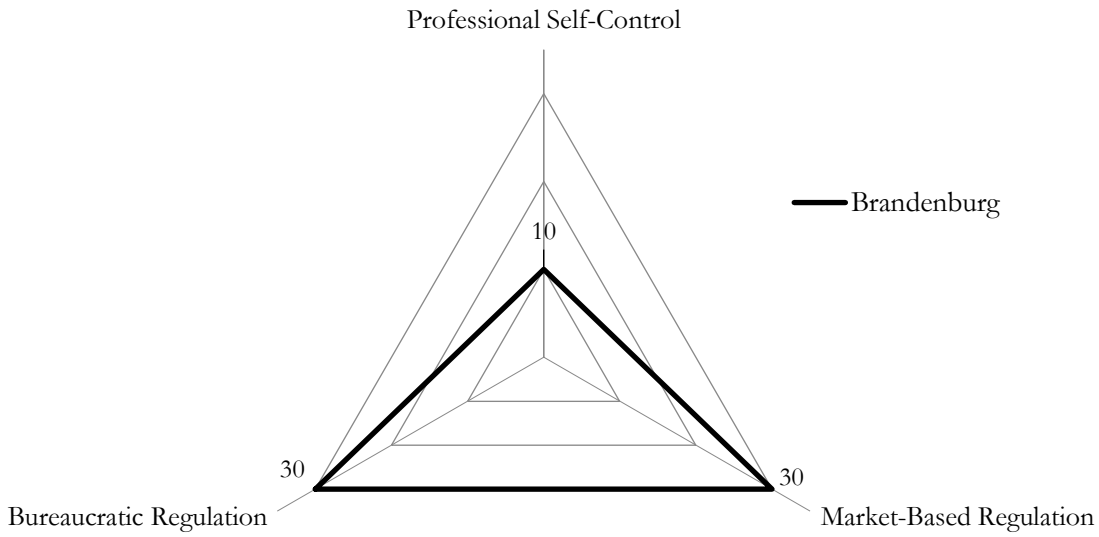

Figure 2. Strength of professional self-control, of bureaucratic regulation and of market-based regulation in the accountability regime of Brandenburg. "10" indicates that the strength of the regulation mode was rated "low", "20" that it was rated "medium" and "30" that it was rated "high".

In contrast to Berlin and Brandenburg, Rhineland-Palatinate places considerable emphasis on data-based bureaucratic regulation and attributes only a minor role to market-based regulation and professional self-control. Subsequent to feedback from external evaluation (like performance tests and school inspections) as well as of internal evaluations, schools are obliged to submit target agreements to the administration. Market-based regulation is rated "low", because schools have some degree of autonomy and school choice is introduced for secondary schools, but evaluation results are not publicly available. Figure 3 presents the strength of the three modes of regulation in Rhineland-Palatinate.

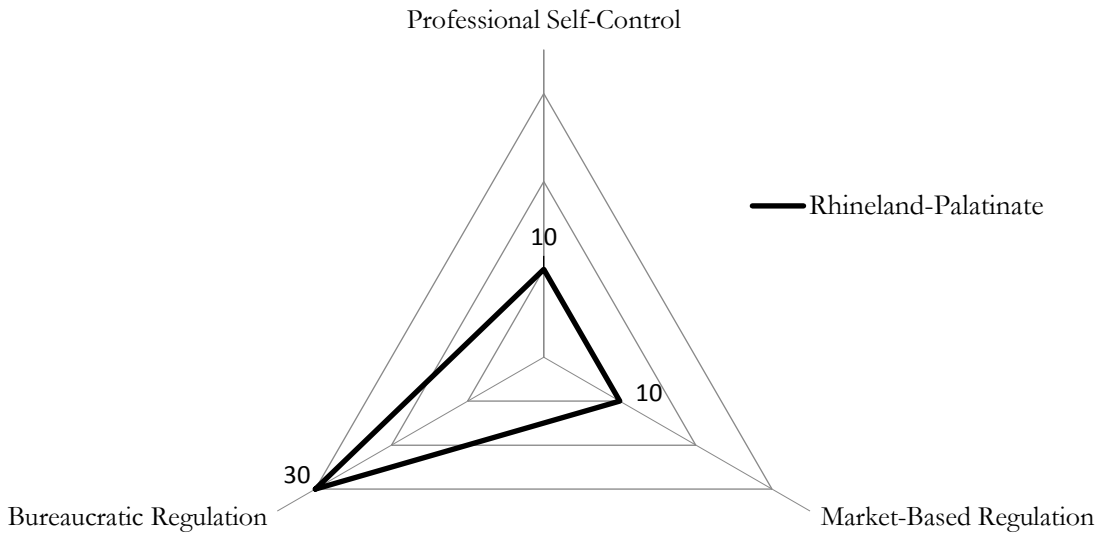

Figure 3. Strength of professional self-control, of bureaucratic regulation and of market-based regulation in the accountability regime of Rhineland-Palatinate. "10" indicates that the strength of the regulation mode was rated "low", "20" that it was rated "medium" and "30" that it was rated "high".

Compared to the other three federal states, Thuringia places the greatest emphasis on professional self-control. The emphasis on professional self-control was rated "medium", because it is left to schools and teachers to deal with evaluation results from performance tests and internal evaluations. The strength of bureaucratic regulation as well as market-based regulation was rated "low", because target agreements between administration and schools are only required subsequent to school inspections. In Thuringia, schools have only a low degree of school autonomy, school choice of secondary schools has not been introduced at a statewide level, and no public data base containing output data has been established. Figure 4 displays the strength of the three regulation modes in Thuringia.

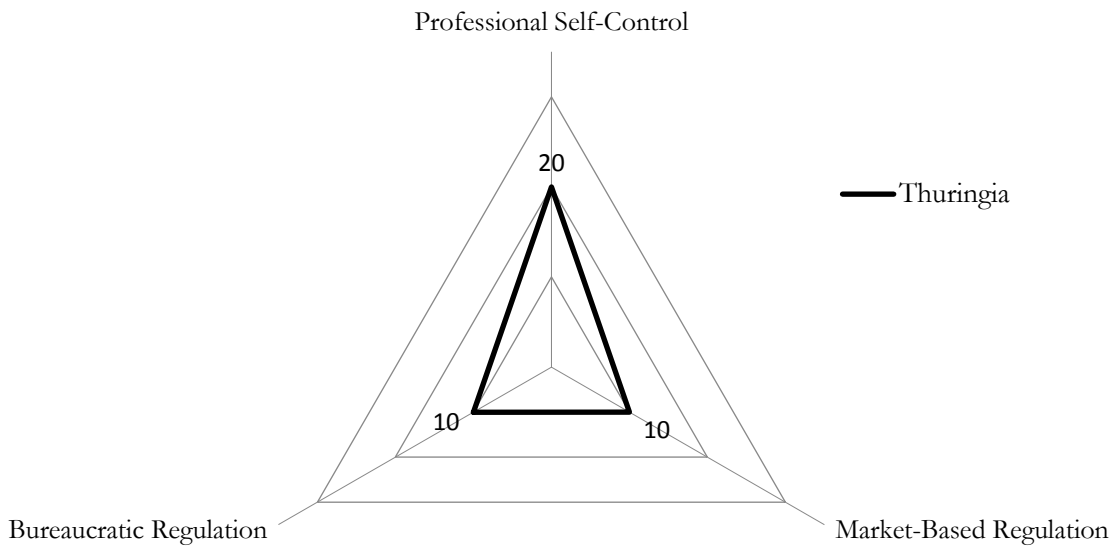

Figure 4. Strength of professional self-control, of bureaucratic regulation and of market-based regulation in the accountability regime of Thuringia. "10" indicates that the strength of the regulation mode was rated "low", "20" that it was rated "medium" and "30" that it was rated "high".

The analysis of policy documents shows that the accountability regimes of the four states can be characterized as either no- or low-stakes contexts. On a scale measuring the degree to which gratifications and sanctions are linked to evaluation results from "no stakes" to "high stakes" the accountability regimes of Berlin and Brandenburg can be characterized as low-stakes rather than no-stakes contexts. In Berlin and Brandenburg, it is compulsory to publish evaluation data, so school choice can be based on a school's achievement in performance tests and school inspections. Thus, parents' choices of schools can function as more informal sanctions or gratifications attached to evaluation results. In addition, formal sanctions are linked to low performances in school inspections in Berlin and Brandenburg. In both states, the period between inspections is shortened if schools are judged as being in need of substantial development, and these schools are put under closer administrative control. In contrast to Berlin and Brandenburg, no formal sanctions and gratifications are bound to evaluation results in Rhineland-Palatinate and Thuringia. Therefore, both states are no-stakes rather than low-stakes contexts, although evaluation data are subject to target agreements between schools and administrators which means, strictly speaking, that evaluation results do not serve as "feedback only". Nevertheless, in comparison to the other three states, the accountability regime of Thuringia comes closest to "no stakes" since least emphasis is put on data-based administrative control.[3]

[3] After the Nefo study terminated in 2014, substantial changes have occurred in the structure of the accountability regime of Thuringia. Since 2015, Thuringia's school inspection program is temporarily suspended. As target agreements were linked to school inspections, data-based administrative control of schools does not play any role anymore. At present, evaluation data do exclusively serve as feedback for schools and teachers.

A List of Side Effects

As outlined above, the survey in the Nefo project was based on a list containing side effects known from the Anglo-American discussion and those detected in the interview study. This list is presented in Table 4. It aims to provide a broad overview of the range of identified side effects, but it does not claim to be exhaustive.

Table 4
List of Side Effects

I. Side effects on the behavior of teachers and principals	
1. On the level of teaching	
Reallocation I	Reallocation of resources between different domains of performance
Reallocation II	Reallocation of resources within certain domains of performance
Reallocation III	Reallocation of individual support
Reallocation IV	Restricting education to performance- and competence considerations
Coaching I	Preparing for formal peculiarities of tests
Coaching II	Preparing for peculiarities of test items with regard to their contents.
Cheating I	Preparing for a certain upcoming test ('teaching to the test')
Cheating II	Support and corrections during the test
Cheating III	Subsequent corrections of test items
Cheating IV	Inappropriate grading to improve results
Myopia	Focusing on short-term external requirements
Ossification	Sticking to routinized methods
Assimilation of teaching I	Adopting a psychometrical test culture for the purpose of teaching
Assimilation of teaching II	Focusing on competence indicators instead of long-term educational processes (orientation towards output instead of orientation towards outcome)
Assimilation of teaching III	Lowering the attainable level of instruction by adjusting it to the level of regulated standards

Table 4 (Cont'd.)
List of Side Effects

2. On the level of schools	
Cream skimming I	Recruiting high performance pupils
Cream skimming II	Expelling weak performers
Predatory competition instead of quality-based competition	Crowding out schools by poaching pupils
Test pool selection I	Occasionally excluding weak performers from tests
Test pool selection II	Permanently excluding weak performers from tests
Test pool selection III	Additional training and delayed testing of pupils with weak performance in preparation for tests
Inspection pool selection	Occasionally excluding poorly performing teachers from school inspection
Mobilization of external resources I	By parents
Mobilization of external resources II	By private service providers
Mobilization of external resources III	By the school environment e.g. companies that support the school
Increase of transaction costs I	By documenting a system of accountability (Shifting resources away from the core business of instruction)
Increase of transaction costs II	By increasing advertising (Shifting resources away from the core business of instruction)
Declining quality by shifting resources	Declining quality of teaching by hiring less qualified staff
Mimetic isomorphism	Adopting 'legitimate models' at the level of instruction and organization
Window dressing	Optimizing the appearance of the school in short-term preparation for school inspections
Work to rule	Restricting activities to the level of set standards
Hierarchization of the principal-teacher-relationship	Monitoring and pressure tend to dominate the principal's role

Table 4 (Cont'd.)
List of Side Effects

II. Side effects impacting on the attitudes of teachers and principals	
Effects of deprofessionalization I	Partly delegating teachers' tasks to experts who are external to the profession
Effects of deprofessionalization II	Allocating tasks to teachers who are external to the profession
Conflicts with professional integrity	Conflicts between required demands and professional educational convictions
Dependency on expert judgements	Feedback from experts becomes the main criterion for self-assessment of teachers and their teaching
Competitive thinking	Teachers' competitive thinking interferes with inner-school cooperation
Change of motivation	Increase or decrease of existing motivations through incentives
Restricted autonomy by accountability	Control is considered to interfere with professional autonomy
Erosion of trust I	Vote of no confidence by educational policy and administration
Erosion of trust II	Distrusting educational policy and administration
Erosion of trust III	Parental distrust
Erosion of trust IV	Distrusting parents and the school environment
Mental costs	Experiencing overextension, stress, fear, frustration, uncertainty etc.

Section I of the table contains side effects at the behavioral level of teachers and principals. This section is divided into behavioral effects at the teaching level and effects at the school level. Section II of the table contains the side effects impacting on teachers' attitudes. Unfortunately, it is beyond the scope of this paper to outline the argument underlying the classification of each of the listed effects, but most of the side effects listed in Table 1 are well known in the Anglo-American debate and therefore need no further explication (Bagley, 2006; Ball, 2003a, 2003b; Finnigan & Gross, 2007; Hamilton et al., 2007; Hargreaves & Goodson, 2006; Hursh, 2007; Koretz, 2005; Nichols & Berliner, 2007; Popham, 2006; Smith, 1993, 1995; Troman, 2000; Valli & Buese, 2007). At this point, we must confine ourselves to elaborating on two phenomena that have been less prominent until now, namely the "assimilation of teaching" and "dependency on expert judgement". The reasons why these effects have not yet received much attention may be that they are examples of

side effects that occur when teachers conform to the measurement system and that both phenomena can be described as the corruption of educators. Until now, a focus on side effects that occur as a result of avoiding accountability pressures and a focus on such side effects that can be understood as corruption of indicators seems to be prevailing in the discussion.

"Assimilation of teaching" entails long-term cultural effects on the conduct of teaching. We found three different forms: (1) Assimilation of teaching is the case with teachers who use test items to teach pupils (Assimilation of teaching I). Proponents of accountability in education might consider this a positive effect. In Germany, the Institute for Educational Quality Improvement (IQB) maintains a database containing test items that were used in former standards-based tests. The database not only contains test items and information on rating answers, but also didactical comments for teachers. Notwithstanding, using test items for teaching can be problematic, because test items and teaching activities fulfill different requirements. While test items aim at evaluating pupils' competencies, didactical tasks are supposed to initiate learning. Thus, the requirements for didactical tasks and test items are inherently different (Benner, 2007; Blömeke, 2009). (2) Another form of assimilation of teaching is the alignment of teaching to competence indicators (output) instead of orienting teaching to long-term developments (outcome). In the literature, various conceptualizations of output and outcome can be found. Here, we follow Chapman (2008) and Schreyer (2002) in defining output as the direct and immediate effects of teaching, i.e. student performance in terms of skills and competencies. Outcome, on the other hand, refers to the indirect long-term effects of schooling, e.g. participation in society (Chapman, 2008; Schreyer, 2002). The greater the gap between output and outcome, the greater the risk of tests being mistaken for indicators of outcome, instead of output (Shavelson & Huang, 2003). In education, measurable output, such as reading literacy, refers to basic skills. Basic skills are no doubt important, but they are relevant because they are fundamental to higher aims, such as educating students to think critically (Wagner, 1989). (3) When analyzing the interviews, we also found that teachers lowered the aspiration level of their teaching in reaction to standards or tests. If this is the case, a class will not achieve its full potential performance level. It is obvious that this substantially contradicts aims at enhancing quality in education.

Similar to different forms of "assimilation of teaching", "dependency on expert judgement" is a side effect of accountability in education which mostly goes unnoticed. As Biesta (2015) points out, teaching is a practice which cannot be guided by predefined aims and clear-cut rules of action. It is a practice which needs to take into account the distinctive features of a pupil or a class. Decisions have to be made in accordance with the characteristics of individual cases. If teachers underestimate the role of judgement in favor of instructive prescriptions, Schütze (1996) speaks of a 'professional self-misunderstanding' (p. 237). He stresses that understanding professional practice as a technological production process results in inefficiency.

Both the "assimilation of teaching" and "dependency on expert judgement" are examples of side effects that become evident only if one acknowledges that not only evading but also adapting to accountability structures constitute problematic effects. The rather explorative comparative, quantitative study of the Nefo study gives hints for the explanation of both adaptive behavior according to incentives and evasive behavior contrary to incentives.

Side Effects as Adaptive Behavior (According to Incentives) and Evasive Behavior (Contrary to Incentives)

The quantitative data analysis of our survey revealed that some side effects are commonly observed by school teachers, while others are not. Accordingly, side effects can be grouped broadly

in two clusters, one containing side effects which are perceived by the majority of teachers in all four states, and the second cluster comprising side effects which are reported by the minority of teachers in at least one federal state. Asking what the side effects that were observed by most respondents have in common, we found that such side effects can be understood as adaptive behavior (according to incentives), while side effects, which were reported by the minority of teachers in at least one federal state, are forms of evasive behavior (contrary to incentives).

It is beyond the scope of this paper to present results for the variety of side effects studied, so we restrict ourselves to the presentation of results for a selection of side effects. Table 5 gives an overview of the frequency distribution of the side effects described in the paper.

Table 5
Response Rates for Selected Items (cf. text)

		Responses				
Side effect	State	Agree completely	Agree partly	Disagree partly	Disagree completely	*N*
Coaching	BE	25.1 (0.04)	41.5 (0.04)	19.1 (0.03)	14.3 (0.03)	229
Do you use test items of older test booklets in preparation for VERA?	RP	18.5 (0.03)	40.7 (0.03)	17.8 (0.02)	23.0 (0.03)	498
	TH	26.2 (0.03)	42.5 (0.03)	21.2 (0.03)	10.1 (0.02)	341
	BB*	39.7 (0.03)	42.6 (0.03)	9.2 (0.02)	8.4 (0.02)	325
Myopia	BE*	27.6 (0.03)	48.0 (0.03)	22.6 (0.02)	1.9 (0.01)	464
I have the impression that education has been guided by the short-term fulfillment of predefined quality measures for the past years.	RP*	15.8 (0.01)	54.0 (0.02)	26.4 (0.02)	3.8 (0.01)	923
	TH	11.7 (0.02)	44.1 (0.03)	38.9 (0.03)	5.3 (0.02)	460
	BB*	18.7 (0.02)	52.3 (0.03)	25.0 (0.02)	4.0 (0.01)	532
Assimilation of Teaching	BE*	20.9 (0.02)	46.4 (0.03)	30.3 (0.02)	2.4 (0.01)	466
I have the impression that the orientation towards long-term educational processes has taken a back seat in recent years.	RP	13.4 (0.01)	42.5 (0.02)	39.2 (0.02)	4.8 (0.01)	933
	TH	11.9 (0.02)	40.4 (0.03)	43.1 (0.03)	4.6 (0.01)	466
	BB*	17.1 (0.02)	48.8 (0.03)	30.3 (0.03)	3.8 (0.01)	528

Table 5 (Cont'd.)
Response Rates for Selected Items (cf. text)

Side effect	State	Responses: Agree completely	Agree partly	Disagree partly	Disagree completely	*N*
Cheating	BE*	7.0 (0.02)	40.9 (0.03)	28.4 (0.04)	23.7 (0.04)	206
Individual teachers use the time between receiving the test booklets and conducting the test to practice some of the upcoming test items with the students.	RP*	7.2 (0.02)	37.5 (0.03)	33.3 (0.03)	22.2 (0.02)	468
	TH	1.2 (0.01)	38.1 (0.03)	30.9 (0.04)	29.8 (0.03)	292
	BB	8.9 (0.02)	39.8 (0.04)	26.2 (0.03)	25.1 (0.04)	292
Cream Skimming	BE	5.3 (0.05)	21.1 (0.09)	31.6 (0.11)	42.1 (0.11)	19
Over the past few years, our school has increased its efforts to avoid accepting pupils with poor performance	RP	9.6 (0.04)	5.8 (0.03)	25.0 (0.06)	59.6 (0.07)	52
	TH	6.2 (0.04)	0.0 (0.0)	18.8 (0.07)	75.0 (0.08)	32
	BB	6.5 (0.04)	9.7 (0.05)	22.6 (0.08)	61.3 (0.09)	31

Note. BE = Berlin; RP = Rhineland-Palatinate; TH = Thuringia; BB = Brandenburg; Information in percent; ( ) Standard errors in parenthesis; a weighted distribution is presented for "Coaching", "Myopia", "Assimilation of Teaching" and "Cheating"; an unweighted distribution is presented in the case of "Cream Skimming"; * Statistically significant differences compared to Thuringia ($p \leq 0.05$); for significance testing, the response categories "agree completely" and "agree partly" were combined to one category "agreement" and the response categories "disagree partly" and "disagree completely"" were combined to one category 'disagreement'.

Adaptive Behavior According to Incentives

Adaptive behavior refers to side effects that might seem unproblematic at first sight because accountability itself serves as reference point for teaching practice. There are two ways in which accountability stimulates adaptive behavior, one way is that certain forms of behavior are conveyed symbolically at the model level, as is the case with Etzioni's observation that what is measured is what matters (1964). Such side effects are essentially systematic effects of accountability of education. The second way is less subtle; certain kinds of behavior are communicated to teachers as desired behavior at the level of accountability regimes. This is achieved, for example, through the decision to establish a database with old test items open to teachers, and is therefore a matter of implementation features.

"Myopia" and "assimilation of teaching" (see Table 4) are examples of the first case. In our survey, we asked teachers to agree or disagree with the statement "I have the impression that education has been guided by the short-term fulfillment of predefined quality measures for the past few years." 76% of teachers in Berlin, 71% in Brandenburg, 70% in Rhineland-Palatinate and 56% in

Thuringia agreed with this statement.[4] At the same time, the majority of teachers in the four states reported that the orientation towards long-term educational processes has taken a back seat in recent years.

Other kinds of adaptive behavior can be understood as the teachers' reaction to opportunities that are created actively by policy makers and administrators. "Coaching" with test items from recent tests is an example of this. Such coaching is actively promoted in Germany at the system level. Not only are old test items made publicly available via the IQB-data base mentioned above, but also the use of the data base is actively suggested in official administrative instructions (Senatsverwaltung für Bildung, Jugend und Wissenschaft, 2013), ministerial circulars (MBJS, 2014) as well as some institutional websites, e.g. Berlin-Brandenburg Institute for School Quality Improvement (ISQ). In all four federal states, the majority of teachers report preparing for standards-based tests with old test items. Eighty-two percent of teachers in Brandenburg report using older versions of test booklets to prepare for tests. Fewer teachers report doing so in Thuringia (69%), Berlin (67%) and Rhineland-Palatinate (59%). A discussion of the positive and negative consequences of coaching for the validity of test results is given by Koretz (2005). Substantive coaching that capitalizes on certain aspects of test content may lead to score inflation, while non-substantive coaching, focusing on arbitrary aspects of test items, may improve validity by removing irrelevant barriers to performance.

Evasive Behavior Contrary to Incentives

While adaptive behavior was reported by the majority of teachers in our study, only the minority of teachers reported such side effects which we termed evasive behavior. Examples at the classroom level include violations of regulations in the context of standards-based tests and school inspections, as well as coaching in terms of using special preparation booklets, which goes beyond actively promoted coaching, as in using old test items to prepare students for tests. At the school level, strategies such as "cream skimming", "test pool selection" and the "exclusion of teachers from school inspections" could be attested. In these cases, teachers try to evade accountability pressures with strategies that are by no means imposed or desired by proponents of accountability.

The degree to which teachers agree with items which refer to forms of evasive behavior, ranges from relatively high agreement to almost no agreement. For instance, teachers agree with the following item: "Individual teachers use the time between receiving the test booklets and conducting the test to practice some of the upcoming test items with the students". 49% of teachers in Brandenburg, 48% in Berlin, 45% in Rhineland-Palatinate and 39% in Thuringia agree with the item. These forms of "cheating" call the comparability of test results into question, but might promise teachers a reasonable way to avoid being blamed for bad test results. With respect to this result, we must concede that the test item is rather complex, so that it might be misinterpreted as asking about "test preparation" rather than "cheating".

Less agreement was reported on the item "Over the past few years, our school has increased its efforts to avoid accepting pupils with poor performance." This item was only administered to school principals, 24% in Berlin, 16% in Brandenburg, 15% in Rhineland-Palatinate and 7% of principals in Thuringia agreed with the item. Thus, "cream skimming" does not seem to be widely spread in the four German federal states, yet it is present even in the context of no and low stakes.

Almost no agreement was found concerning the item "Sometimes, individual teachers are encouraged to stay home on days when the school inspection takes place." Even though almost no teachers agreed with the item, it is still remarkable that – contrary to what we expected – at least a

[4] "Agreement" comprises the categories "agree completely" and "agree partly".

small number of teachers in Berlin, Rhineland-Palatinate and Brandenburg did report that individual teachers are advised to stay home while a school inspection is conducted. In Thuringia, none of the teachers agreed with the item.

When we look at the response pattern to the variety of items operationalizing evasive behavior, it is evident that in almost all cases, teachers in Thuringia report less agreement than teachers in Berlin, Brandenburg and Rhineland-Palatinate. Assuming that the reported agreement indicates the actual occurrence of side effects, we conclude that evasive behavior is less widespread in Thuringia than in the other three states. As stated above, the comparative design of the study suggests interpreting this pattern as an effect of the emphasis and interplay between the regulation modes of accountability in education, which is specific to each federal state. Thus, it can be argued that the accountability-structures in Thuringia stimulate evasive behavior to a lesser degree than the accountability-structures of the other three states. In contrast to these states, competition-based and bureaucratic pressures play only a minor role in Thuringia. Here, data-based quality management of schools is left largely to teachers.

Conclusions and Outlook

The findings of the Nefo project reveal that side effects are not bound to high-stakes contexts, but that they are also a substantial problem in low- or no-stakes contexts. Notwithstanding, a differentiation between adaptive behavior according to incentives, and evasive behavior contrary to incentives, must be made. While adaptive behavior is observed by the majority of teachers and principals in the no- and low-stakes contexts of Berlin, Brandenburg, Rhineland-Palatinate and Thuringia, evasive behavior is not.

Adaptive behavior may result from the alignment of teaching either to the ideal-type accountability model or to specific characteristics of accountability regimes. In the first case, side effects are a systematic factor of accountability in education. These side effects occur almost by default with the implementation of accountability in education. In the second case, adaptive behavior can be understood as an effect of implementation features. The extent to which these side effects occur could possibly be reduced by changing the implementation characteristics, but one cannot be certain that side effects of this kind would disappear completely after changing the characteristics.

Compared to adaptive behavior, evasive behavior seems to be more strongly connected with implementation characteristics. Accountability regimes with a high degree of market and bureaucratic pressure seem to foster evasive behavior. Even under no- and low-stakes conditions, administrative and market pressures are exerted on schools, since external actors may base their decisions on the results of school evaluations. Administrators may decide to subject a school to closer scrutiny and parents might base their choice of school on such evaluation results. It seems reasonable though, to assume that the implementation of high stakes intensifies these pressures that are inherent to bureaucratic and competition-based regulation.

Further research is needed to test and modify the findings of the Nefo study. A study comparing the distribution of both kinds of side effects — evasive and adaptive behavior — in no- and low-stakes contexts and high-stakes contexts could test two central findings of the study: (1) that adaptive behavior, that is, the assimilation of teaching, is foremost a systematic problem of accountability in education and therefore not necessarily a matter of specific implementation features; and (2) that evasive behavior increases with growing pressure on schools and educators.

Irrespective of how future research might turn out, the Nefo project indicates that a broadened view of side effects can yield a deeper understanding of the problematic effects of accountability in education. However, in order to reach that point, some well-established truths will

have to be called into question. This is true both for research on accountability in education in high-stakes contexts, and for research on educational accountability in low-stakes or no-stakes contexts. With regard to the former line of research, evidence from the Nefo project could problematize a widespread one-sided concentration on high stakes as the alleged major problem of accountability in education. Although it is understandable that high stakes have become a primary concern for both educators and researchers in many English-speaking countries, it might be important to note that high stakes, as a contingent implementation characteristic of accountability in education, are not the only problem in the context of side effects. Countries with a low-stakes or no-stakes version of accountability in education should by no means be treated as if they were the promised land of education policy.

With regard to the latter line of research, evidence from the Nefo project could also problematize a widespread one-sided concentration on allegedly inadequate reactions of teachers to accountability instruments. When teachers do use data from standards-based tests and school inspections, as they are supposed to, but in fact produce a wide range of side effects even in contexts of no stakes or low stakes, the common practice of continually blaming teachers for their incompetence or resistance to instruments that are flagged as "feedback only", is barely convincing. Furthermore, countries with a high-stakes version of accountability in education should not serve as a "cautionary tale" to eliminate concerns about the side effects of accountability in education in general. Both lines of research can use evidence from the Nefo project as an opportunity to investigate the side effects of accountability in education as systematic effects of indicator-based regimes of quality assessment and assurance — irrespective of whether or not there are really high stakes.

Acknowledgment

This work is based on the research project "Unintended Effects of Accountability in the School System" supported by the Federal Ministry of Education and Research, Germany under Grant No. 01JG1006.

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. https://doi.org/10.1207/s15430421tip4201_4

Amrein, A. L., & Berliner, D. C. (2002). *An analysis of some unintended and negative consequences of high-stakes testing*. Retrieved from http://nepc.colorado.edu/files/EPSL-0211-125-EPRU.pdf

Bagley, C. (2006). School choice and competition: A public-market in education revisited. *Oxford Review of Education, 32*(3), 347-362. https://doi.org/10.1080/03054980600775656

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., … Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (Economic Policy Institute Briefing Paper No. 278). Retrieved from http://files.eric.ed.gov/fulltext/ED516803.pdf

Ball, S. J. (2003a). Market mixes, ethical re-tooling and consumer heroes: Education markets in England. In M. Mangold & J. Oelkers (Eds.), *Demokratie, Bildung und Markt* (pp. 257-279). Bern: Peter Lang.

Ball, S. J. (2003b). The teacher's soul and the terrors of performativity. In S. J. Ball (Ed.), *Education Policy and Social Class* (pp. 143-156). London: Routledge.

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 [Side effects of accountability in "no-" and "low-stakes" contexts. Qualitative and quantitative findings of a study in four federal states]. In Federal Ministry of Education and Research (Ed.), *Steuerung im Bildungssystem. Implementation und Wirkung neuer Steuerungsinstrumente im Schulwesen* (pp. 208-238). Bielefeld: W. Bertelsmann.

Bellmann, J., & Weiß, M. (2009). Risiken und Nebenwirkungen Neuer Steuerung im Schulsystem. Theoretische Konzeptualisierung und Erklärungsmodelle. [Risks and side effects of accountability in the school system: Theoretical conceptualization and explanatory models]. *Zeitschrift für Pädagogik, 55*(2), 286-308. Retrieved from http://www.pedocs.de/volltexte/2011/4251/pdf/ZfPaed_2009_2_Bellmann_Weiss_Risiken_Neue_Steuerung_D_A.pdf

Benner, D. (2007). Unterricht – Wissen – Kompetenz. Zur Differenz zwischen didaktischen Aufgaben und Testaufgaben [Instruction – knowledge – competence. On the difference between didactical tasks and test items]. In D. Benner (Ed.), *Bildungsstandards. Instrumente zur Qualitätssicherung im Bildungswesen. Chancen und Grenzen – Beispiele und Perspektiven* (pp. 123-138). Paderborn: Ferdinand Schöningh.

Betebenner, D. W., Howe, K. R., & Foster, S. S. (2005). On school choice and test-based accountability. *Education Policy Analysis Archives, 13*(41), 1-19. https://doi.org/10.14507/epaa.v13n41.2005

Biesta, G. (2015). What is education for? On good education, teacher judgement and educational professionalism. *European Journal of Education, 50*(1), 75-87. https://doi.org/10.1111/ejed.12109

Blömeke, S. (2004). Empirische Befunde zur Wirksamkeit der Lehrerbildung [Empirical findings on the efficacy of teacher education]. In S. Blömeke, P. Reinhold, G. Tulodziecki & J. Wildt (Eds.), *Handbuch Lehrerbildung* (pp. 59-91). Bad Heilbrunn: Klinkhardt.

Blömeke, S. (2009). Allgemeine Didaktik ohne empirische Lernforschung. Perspektiven einer reflexiven Bildungsforschung [General didactics without empirical learning research. Perspectives of reflexive education research]. In K.-H. Arnold, S. Blömeke, R. Messner & J. Schlömerkemper (Eds.), *Allgemeine Didaktik und Lehr-Lernforschung. Kontroversen und Entwicklungsperspektiven einer Wissenschaft vom Unterricht* (pp. 13-27). Bad Heilbrunn: Klinkhardt.

Böttcher, W. (2012). Teaching to the Test. Warnung vor dem falschen Vorbild [Teaching to the test. Warning about the bad example]. *Friedrich Jahresheft, 30*, 88–89.

Campbell, D. T. (1975). Assessing the impact of planned social change. In G. M. Lyons (Ed.), *Social research and public policies* (pp. 3–45). The Dartmouth/OECD Conference. Hanover: Dartmouth College.

Chapman, D. W. (2008). Options for improving the management of education systems. In W. K. Cummings & J. H. Williams (Eds.), *Policy-making for education reform in developing countries: Policy, options and strategy* (pp. 251-276). Lanham: Rowman & Littlefield Education.

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

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18*(1), 105-115. https://doi.org/10.1037/h0030644

DfE (Department for Education). (2010). *The importance of teaching. The schools white paper 2010*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/175429/CM-7980.pdf

Döbert, H., Rürup, M., & Dedering. K. (2008). Externe Evaluation von Schulen in Deutschland – die Konzepte der Bundesländer, ihre Gemeinsamkeiten und Unterschiede [External evaluation of schools in Germany – the concepts of the federal states, their similarities and differences]. In H. Döbert & K. Dedering (Eds.), *Externe Evaluation von Schulen* (63-151). Münster: Waxmann.

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. https://doi.org/10.1111/j.1467-8527.2006.00333.x

Etzioni, A. (1964). *Modern organizations*. Englewood-Cliffs: Prentice-Hall.

Finnigan, K. S., & Gross, B. (2007). Do accountability policy sanctions influence teacher motivation? Lessons from Chicago's low-performing schools. *American Educational Research Journal, 44*(3), 594-629. https://doi.org/10.3102/0002831207306767

Frey, B. S. (2012). Crowding out and crowding in of intrinsic preferences. In E. Brousseau, T. Dedeurwaerdere & B. Siebenhüner (Eds.), *Reflexive governance of global public goods* (pp. 75-83). Massachusetts Institute of Technology. https://doi.org/10.7551/mitpress/9780262017244.003.0087

Frey, B. S., & Osterloh, M. (2006). *Evaluations: hidden costs, questionable benefits, and superior alternatives* (University of Zurich IEW Working Paper No. 302). Retrieved from Social Science Research Network website: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=928354

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J. L., … Barney, H. (2007). *Standards-based accountability under No Child Left Behind. Experiences of teachers and administrators in three states*. Santa Monica: Rand Corporation.

Hargreaves, A., & Goodson, I. (2006). Educational change over time? The sustainability and nonsustainability of three decades of secondary school change and continuity. *Educational Administration Quarterly, 42*(1), 3-41. https://doi.org/10.1177/0013161X05277975

Hopmann, S. T. (2003). On the evaluation of curriculum reforms. *Journal of Curriculum Studies, 35*(4), 459-478. https://doi.org/10.1080/00220270305520

Hout, M., & Elliott, S. (Eds.). (2011). *Incentives and test-based accountability in education*. Washington: The National Academies Press.

Hursh, D. (2007). Assessing No Child Left Behind and the rise of neoliberal education policies. *American Educational Research, 44*(3), 493-518. https://doi.org/10.3102/0002831207306764

Jäger, D. (2012). Herausforderung Zentralabitur: Unterrichtsinhalte variieren und an Prüfungsthemen anpassen [Challenge central exams: Varying the curriculum in alignment to test contents]. In K. Maag Merki (Ed.), *Zentralabitur. Die längsschnittliche Analyse der Wirkungen der Einführung zentraler Abiturprüfungen in Deutschland* (pp. 179–205). Wiesbaden: VS. https://doi.org/10.1007/978-3-531-94023-6_8

Jones, G., Jones, B., & Hargrove, T. (2003). *The unintended consequences of high-stakes testing*. Lanham: Rowman & Littlefield.

Koretz, D. M. (2005). Alignment, high stakes, and the inflation of test scores. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (pp. 99–118). Malden: Blackwell. https://doi.org/10.1111/j.1744-7984.2005.00027.x

Levin, H. M. (1998). Educational vouchers: Effectiveness, choice, and costs. *Journal of Policy Analysis and Management, 17*(3), 373-392. https://doi.org/10.1002/(SICI)1520-6688(199822)17:3<373::AID-PAM1>3.0.CO;2-D

McLaughlin, P. (2002). Functional explanation. In R. Mayntz (Ed.), *Akteure – Mechanismen Modelle. Zur Theoriefähigkeit makro-sozialer Analysen* (pp. 196-212). Frankfurt: Campus Verlag.

Ministerium für Bildung, Jugend und Sport des Landes Brandenburg. (2014, September). *Rundschreiben Nr.10/2014 vom 15.09.2014 [Circular Number 10/2014 September 15th 2014].* Retrieved from https://www.isq-bb.de/uploads/media/Rundschreiben_2015_VERA-3_8_2014-09-15.pdf

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

Moe, T. M. (1995). Private vouchers. In T. M. Moe (Ed.), *Private vouchers* (pp. 1-40). Stanford: Hover Institution Press.

Nichols, S., & Berliner, D. (2005). *The inevitable corruption of indicators and educators through high-stakes testing.* [EPSL Report]. Tempe: Arizona State University. http://nepc.colorado.edu/files/EPSL-0503-101-EPRU.pdf

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

Nichols, S., Glass, G., & Berliner. D. (2005). *High-stakes testing and student achievement: Problems for the No Child Left Behind act.* [EPSL Report]. http://files.eric.ed.gov/fulltext/ED531184.pdf

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers*. Retrieved from http://www.bc.edu/research/nbetpp/statements/nbr2.pdf

Popham, J. W. (2006). Educator cheating on No Child Left Behind tests. Can we stop it? *Education Week*, *25*, 32f.

Ravitch, D. (2016). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.

Ridgway, V. F. (1956). Dysfunctional consequences of performance measurements. *Administrative Science Quarterly*, *1*(2), 240-247. https://doi.org/10.2307/2390989

Ryan, K. E., & Feller, I. (2009). Evaluation, accountability, and performance measurement in national education systems. Trends, methods, and issues. In K. E. Ryan & J. B. Cousins (Eds.), *The SAGE international handbook of educational evaluation* (pp. 171-189). Thousand Oaks, CA: SAGE Publications. https://doi.org/10.4135/9781452226606

Schreyer, P. (2012). Output, outcome and quality adjustment in measuring health and education services. *Review of Income and Wealth*, *58*(2), 257-278. https://doi.org/10.1111/j.1475-4991.2012.00504.x

Schütze, F. (1996). Organisationszwänge und hoheitsstaatliche Rahmenbedingungen im Sozialwesen: Ihre Auswirkungen auf die Paradoxien professionellen Handelns [Organizational restrains and governmental conditions in the social sector: Implications for the paradoxes in professional action]. In A. Combe & W. Helsper (Eds.), *Pädagogische Professionalität* (pp. 183-276). Frankfurt am Main: Suhrkamp.

Senatsverwaltung für Bildung, Jugend und Wissenschaft. (2013). *Verwaltungsvorschrift Schule Nr. 11/2013 [Administrative regulation school number 11/2013]*. Retrieved from https://www.isq-bb.de/uploads/media/Verwaltungsvorschrift_11_2013_01.pdf

Shavelson, R. J., & Huang, L. (2003). Responding responsibly to the frenzy to assess learning in higher education. *Change, 35*(1), 10-19. https://doi.org/10.1080/00091380309604739

Smith, P. (1993). Outcome-related performance indicators and organizational control in the public service sector. *British Journal of Management 4*: 135-151. https://doi.org/10.1111/j.1467-8551.1993.tb00054.x

Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration, 18*(2/3), 277-310. https://doi.org/10.1080/01900699508525011

Thiel, F., Cortina, K. S., & Pant, H. A. (2014). Steuerung im Bildungssystem im internationalen Vergleich [Governance in the school system. An international comparison]. In R. Fatke & J. Oelkers (Eds.), *Das Selbstverständnis der Erziehungswissenschaft: Geschichte und Gegenwart (Zeitschrift für Pädagogik, Beiheft 60)* (*;* pp. 123-138). Weinheim: Beltz Juventa.

Troman, G. (2000). Teacher stress in the low-trust society. *British Journal of Sociology of Education, 21*(3), 331-353. https://doi.org/10.1080/713655357

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge: Harvard University Press.

Valli, L., & Buese, D. (2007). The changing roles of teachers in an era of high-stakes accountability. *American Educational Research Journal, 44* (3), 519-558. https://doi.org/10.3102/0002831207306859

van Thiel, S., & Leeuw, F. (2002). The performance paradox in the public sector. *Public Performance & Management Reviews, 35*(3), 267-281. https://doi.org/10.1080/15309576.2002.11643661

Vergleichsarbeiten 3. und 8. Jahrgangsstufe (VERA-3 und VERA-8). (n.d.). *Didaktische Handreichung Modul A*. Retrieved from https://www.bildung-lsa.de/files/a70ce5e681388a8af0556178bc76721c/VERA_Did_Hdrg_Modul_A.pdf

Wagner, R. B. (1989). *Accountability in education: A philosophical inquiry*. New York: Routledge.

Wells, A. S., & Holme, J. J. (2005). Marketization in education: Looking back to move forward with a stronger critique. In N. Bascia, A. Cumming, A. Datnow, K. Leithwood & D. Livingstone (Eds.), *International handbook of educational policy* (pp. 19-51). Bodmin: Springer.

Whitty, G., & Power, S. (2000). Marketization and privatization in mass education systems. *International Journal of Educational Development, 20*, 93-107. https://doi.org/10.1016/S0738-0593(99)00061-9

Woessmann, L., & Schuetz, G. (2006). *Efficiency and equity in European education and training systems* (EENEE Analytical Report No. 1). Retrieved from http://www.eenee.de/eeneeHome/EENEE/Analytical-Reports.html

About the Authors

Corrie Thiel
University of Muenster
CorrieThiel@uni-muenster.de
Corrie Thiel is research associate at the Institute of Educational Science at the University of Muenster. In 2017, she completed her doctoral thesis *Between data and cases. Theoretical and empirical analyses of the relation between accountability and professionality in education*. Her research interests include professionality in education, accountability and methodologies in educational research.

Sebastian Schweizer
University of Muenster
Sebastian.Schweizer@uni-muenster.de
Sebastian Schweizer is research associate at the Institute of Educational Science at the University of Muenster. He currently works on his dissertation in which he studies figurations of competition and positioning practices of schools in the school system. He is interested in questions of competition and choice in education, the economization of education, accountability in education as well as of practice theory and discourse analysis.

Johannes Bellmann
University of Muenster
Johannes.Bellmann@uni-muenster.de
Johannes Bellmann is professor at the Institute of Educational Science at University of Muenster. His research interests include educational theory, historical perspectives in educational theory and educational policy. He was coordinator of the project "Unintended Effects of Accountability in the School System" on which the article is based.

About the Guest Editors

Jessica Holloway
Deakin University
Jessica.holloway@deakin.edu.au
Jessica Holloway, Ph.D., is a post-doctoral research fellow at the Centre in Research for Educational Impact (REDI) at Deakin University. She draws on post-structural theory to understand contemporary modes of accountability and its production of new teacher and leader subjectivities. Her current project, entitled *Teacher Leaders and Democracy: An International Study,* looks at modes of distributive leadership in U.S. and Australian schools.

Tore Bernt Sørensen
University of Bristol
t.b.sorensen@bristol.ac.uk
Tore Bernt Sørensen completed his doctorate at the Graduate School of Education, University of Bristol, UK, in 2017 with the dissertation "*Work In Progress: The Political Construction Of The OECD Programme Teaching And Learning International Survey*". Tore's research centres on comparative studies of education governance in a global context. Tore has a background as teacher and teacher trainer in

Denmark. Before starting his doctorate, he worked in the Analysis and Studies Unit of the European Commission's Directorate-General for Education and Culture.

Antoni Verger
Universitat Autònoma de Barcelona
Antoni.verger@uab.cat
Antoni Verger is associate professor at the Department of Sociology of the UAB. A former post-doctoral fellow at the University of Amsterdam, Antoni's research analyses the relationship between global governance institutions and education policy, with a focus on the study of public-private partnerships and accountability policies in education. Currently, he is coordinating the research project REFORMED - *Reforming Schools Globally: A Multiscalar Analysis of Autonomy and Accountability Policies in the Education Sector* (ERC StG, 2016–2021).

SPECIAL ISSUE
Global Perspectives on High-Stakes Teacher Accountability Policies

education policy analysis archives

Volume 25 Number 93 August 21, 2017 ISSN 1068-2341