International Journal of Educational Research Open 3 (2022) 100165

Contents lists available at ScienceDirect

International Journal of Educational Research Open

journal homepage: www.elsevier.com/locate/ijedro

In search of quality and equity: The United Kingdom and Germany in the struggle for PISA scores

Richard Muench [a,b,*], Oliver Wieczorek [a,b]

[a] Senior Professorship of Social Theory and Comparative Macrosociology, Zeppelin University, Am Seemooser Horn 20, Friedrichshafen 88045, Germany
[b] Department of Sociology, especially Sociological Theory, University of Bamberg, Feldkirchenstraße 21, Bamberg 96052, Germany

ARTICLE INFO

Keywords:
Autonomy
Leadership
Accountability
Quality
Equity
Governance

ABSTRACT

School reforms aiming at improving educational performance (quality), closing achievement gaps and reducing the impact of family background on performance (equity) have been on the agenda worldwide for three decades. These reforms converge in a common core which puts emphasis on school autonomy, free school choice, competition between schools, managerial school leadership, high teacher quality and test-based accountability of schools. We investigate how far these reforms have been associated with improvements in quality and equity in two countries following their own developmental path: the United Kingdom and Germany. Despite all insights provided by the existing literature, we still do not sufficiently know how far the application of governance tools following the global reform agenda does make a difference between schools. For closing this research gap, we conduct multi-level linear regression analyses to test the association of governance tools of the reform agenda with individual student performance and achievement gaps based on family background. We make use of data from the Programme for International Student Assessment (PISA) and look at the test years 2000, 2009 and 2015 to see how far there has been improvement over time along with advancing reforms. The results show that the reforms have failed so far.

Introduction and theoretical foundations

Governing the school has widely changed under the influence of a "Global Educational Reform Movement" (GERM) in the past three decades. For improving quality and equity of education, this movement recommends "competition and choice", "standardization", "increased emphasis on reading literacy, mathematics and science," "borrowing of change models from the corporate world" and "test-based accountability policies" (Sahlberg, 2016, pp. 133-136; cf. Sahlberg, 2015; Ball, 2012; Münch, 2020). Improving "quality" means increasing the educational performance of students, improving "equity" means decreasing the impact of family background on educational performance and closing the gap between the top and bottom socioeconomic quarter of students as well as between high and low performers. A major agent of GERM is the OECD with its Programme for International Student Assessment (PISA), conducted every three years with a growing number of participants, ranging from 43 in 2000 to 79 in 2018 (Sjøberg, 2019). Data from PISA are widely used for assessing school systems and recommending reforms, by the OECD (2011, 2015) itself and in publications of education-industrial players such as McKinsey on how schools improve and achieve (Barber & Mourshed, 2007; Mourshed et al., 2010) and Pearson's (2012) *The Learning Curve* (cf. Münch, 2020). The OECD has consecutively focused on school autonomy, free school choice, competition between schools and regular central testing, teacher quality and school leadership (OECD, 2016b, pp. 81–103). As emphasized by the OECD, school autonomy needs to be complemented by managerial school leadership, highly qualified teachers and strict accountability of principals and teachers (OECD, 2016b, p. 114, with reference to Hanushek et al., 2013). According to principal-agent theory, strong managerial school leadership and a strict accountability regime based on large-scale assessments accompanied by decreasing professional autonomy of teachers is expected to prevent schools from moral hazard when granted greater autonomy (cf. Courtney & Gunter 2015; Salokangas & Ainscow, 2017; Verger et al., 2019). Against this backdrop, using variables available in the PISA database, we set up hypothesis H1:

H1: School systems and schools improve performance and minimize performance gaps through reducing the impact of family background by

* Corresponding author at: Senior Professorship of Social Theory and Comparative Macrosociology, Zeppelin University, Am Seemooser Horn 20, Friedrichshafen 88045, Germany.
E-mail address: richard.muench@zu.de (R. Muench).

https://doi.org/10.1016/j.ijedro.2022.100165
Received 24 December 2021; Received in revised form 7 April 2022; Accepted 7 April 2022

applying governance tools. These tools include (1) enlarged autonomy and self-government of schools, complemented by (2) strong school leadership, and (3) regular monitoring of teaching and student performance through tests and assessments.

Family background refers to the socioeconomic status of individual students, the average socioeconomic status of a school's students, student as well as average school migrant background, and test language spoken at home. What cannot be directly included using PISA data is "highly qualified teachers". In this database, these are teachers on ISCED levels 5 or 6, and it is more likely that such teachers are simply employed by schools on higher socioeconomic status and performance than such teachers being the reason for higher performance. However, we may at least assume that strong school leaders and strong accountability measures imply selective recruitment and regular training of teachers, both features of what is considered as enhancing teacher quality by the reform movement. If this is correct, "highly qualified teachers" may be indirectly represented in hypothesis H1.

More recently, the OECD (2016c, 2019b) has focused on school climate. There are various aspects of school climate, and their association with PISA scores is ambivalent. As they are deeply rooted in lifestyles differing according to culture and socioeconomic class, it is very difficult to apply them as governance tools. Certainly, a good school climate is preferrable to a bad climate by principals, teachers, students, and parents alike. Nevertheless, it is much more a feature of schools resulting from the interaction of factors which are not directly under control of principals. Therefore, it is always on the agenda, but cannot directly be taken as a governance tool. One aspect which seems to be clear in its association with performance, is school discipline. Therefore, we account for student and school disciplinary climate separately as control variables.

In the perspective of GERM, the right school governance helps to overcome the impact of family background on student performance (cf. Morgan, 2017). Opposed to this assumption is a theoretical perspective which derives from Bourdieu and Passeron's theory of the indirect school mode of reproducing social inequality based on inherited cultural capital as against the direct family mode based on inherited economic capital (Bourdieu, 1984, 1998; Bourdieu & Passeron, 1990). In this perspective, GERM governance tools do not remove the obstacles for educational achievement residing in disadvantaged family background and lacking cultural capital. They do not eliminate the fundamental disadvantages of low-income families. Therefore, socioeconomically disadvantaged students do not benefit from more autonomous schools, strong principals and frequent tests and assessments. Socioeconomic background remains the central factor associated with educational achievement (cf. Van Zanten, 2005; Lubienski et al., 2021).

In John Goldthorpe's (2007a) view, Bourdieu and Passeron's class theory of education is refuted by the enormous expansion of secondary and tertiary education from the 1960s on and up to now. As he argues, based on Boudon's (1974) distinction between primary and secondary effects of family socialization, compensatory education through schooling can help to correct primary socialization disadvantages and informing parents about the high returns of schooling can help to correct secondary socialization disadvantages. Nevertheless, Goldthorpe (2007b, p. 166) also recognizes that low-income parents are disadvantaged in ensuring the educational achievement of their children (see also Erikson & Goldthorpe, 1992; Breen & Jonsson, 2005; Hertel & Groh-Samberg, 2019). Accordingly, directly opposed to hypothesis H1, we set up hypothesis H2:

H2: As school systems and schools apply governance tools like (1) enlarged autonomy and self-government of schools, complemented by (2) strong school leadership and (3) regular monitoring of teaching and student performance through tests and assessments, these tools do not remove disadvantages of students from low-income families so that they do not change the impact of socioeconomic background on student achievement.

We aim at testing hypotheses H1 and H2 in two steps. In the first step, we carry out a literature review of school governance in two selected countries which have implemented reforms according to GERM on their own developmental path: the United Kingdom as a forerunner of the reform movement with a tradition of liberalism and Germany as a latecomer with a bureaucratic-professional tradition. As the OECD reports on PISA and the corresponding research literature are predominantly focused on comparing countries, there is a lack of knowledge on the direct impact of implementing GERM governance tools on the school level on the performance of individual students. Closing this gap is the aim of the second step of our study. In this step, after a brief look at GERM and PISA performance of a larger number of countries, we turn from the country level to the school and individual student levels. We conduct a multilevel linear regression analysis including school-level and individual-level variables on the association of socioeconomic status, migrant background, disciplinary climate, and governance tools with the variance in individual students' PISA test scores in both countries in 2000, 2009 and 2015.

The reform strategy of comprehensive market-based monitoring of schools in the UK

The education system in the United Kingdom was characterized by its local administration until the 1980s. The individual schools had a relatively high degree of autonomy. The Local Education Authorities (LEAs) were the main control bodies. At the end of the 1980s, education policy was re-directed, most of all in England. In all four jurisdictions of the UK reforms have taken place, following the leadership of the UK government. Throughout the UK, measures of increasing accountability of schools have been implemented. Nevertheless, there are differences in the interpretation of these measures between the four jurisdictions. Scotland (and less consequently Wales) did not follow England's lead in marketization reforms. In the following, the outline of reforms and related research is focused on England (and Wales to some part) (cf. Croxford & Raffe, 2007; Raffe et al., 1999; Reynolds & Mckimm, 2020; Woods et al., 2021). In order to improve the quality of school education and minimize its costs, the emphasis was on the introduction of market forces and competition between schools for pupils and resources. At the same time, control over school performance was extended and centralized with the national government (Ball, 2008). The result was to be a "self-improving school-led system" (Greany & Higham, 2018). In the wake of this reform movement, the reference to international performance comparisons and the front-runners of these comparisons in East Asia has become increasingly important (You, 2017).

The Educational Reform Act of 1988 (OPSI, 1989) was one of the first measures of the new governance of the education system. The schools were given greater decision-making leeway in management and budgeting. At the same time, their previously existing autonomy in curriculum design was restricted by the introduction of a national curriculum. At the heart of the reform measures was the establishment of a quasi-market in the education system. Public and private schools are to compete for pupils and resources (Whitty et al., 1998). This is the line taken by the flat-rate student allowance introduced in 1988. According to this device, the more students the schools can attract, the more money they receive. On the parents' side, this includes the free choice of school. In this context, the options have been expanded by creating different types of schools (Angus, 2015). The idea is that the basis for parents' free choice of school should be annual league tables of schools.

In July 2010, the British parliament passed a law that allows all schools, primary, secondary and special schools, to become so-called "academies" (Heilbronn, 2016). Academies are publicly funded independent schools that enjoy greater autonomy than conventional schools. However, it is not only primary and secondary schools that are

encouraged to obtain the status of academies. In addition, the "Free Schools" have created the opportunity for parents, teachers, universities or companies to establish their own schools oriented towards local problems which also receive state funding (cf. Gunter & McGinity, 2014; Rayner et al., 2018).

In return for increasing school autonomy and free choice of school, central control over school performance and teaching was expanded. The Teacher Training Agency (TTA) was created in 1994 for the initial and in-service training of teachers and the monitoring of their performance, and in 2005 it was transformed into the Training and Development Agency for Schools (TDA) (Ellis et al., 2019). In April 2012, this task was transferred to the newly formed Teaching Agency (TA) with a broader range of duties. Performance-related pay and, since 1998, Advanced Skills Teachers (AST) in England and Wales have been responsible for ensuring that teachers meet the performance standards set. The AST program was practiced until August 2013.

The quality of the schools themselves is regularly checked both by the school management itself in self-evaluation and by independent school inspectorates. The control regime has been internalized by school directors and teachers to such an extent that one can speak of a system of post-panoptism (see Perryman, 2006; Perryman et al., 2018; Ball et al., 2012; Courtney, 2016). Principals must be very inventive to succeed "against" this system (Greany & Waterhouse, 2016). School Effectiveness Research (SER), which was introduced by the Thatcher government in the 1980s and continued by New Labour under Tony Blair and Secretary of State for Education David Blunket, has gained great importance for school policy (Reynolds, 2010; Perry, 2017).

What is the evidence regarding the outcome of these reforms and regarding hypotheses H1 and H2? Stephen Ball (1993) warned in the early stages of the reforms that the free choice of school in education markets leads to greater social inequality. Early evidence proved that middle-class parents benefit from free school choice, but not working-class parents (Ball et al., 1996). Gorard (2014) reports that newly established academies are no better than schools they have replaced or are in competition with, and that they reproduce rather than overcome the existing segregation of neighborhoods as intended. In his already classic study of the outcome of free school choice in South Wales, Gorard (2019) found that it was not associated with rising student achievement levels or decreasing social stratification. However, Eyles et al. (2016) established that students in disadvantaged metropolitan areas, in particular, are more likely to graduate from school and obtain a university degree when their school was converted to an academy than a control group of students attending a school that was not turned into an academy. However, there is enough evidence to support the objection that this may be true for one school or another with a well-selected, particularly motivated student body, but not generally. In particular, the authors neglect the fact that the competition between schools, which has been intensified by the academy program, is primarily related to the recruitment of the best performing pupils in a neighborhood, and therefore leads to growing inequality in the educational performance of schools in this neighborhood. For example, Allen and Higham (2018) examined 325 new free schools, founded between 2011/12 and 2015/16, and discovered that they enroll a socioeconomically better off student body than their neighborhood. Accordingly, free school choice ultimately reinforces social polarization (Greany & Higham, 2018; Burgess et al., 2019; Hansen & Gustafsson, 2019; Münch, 2020).

In addition to the increase in educational inequality, there is growing complaint about the subjugation of schools to a comprehensive control regime. The establishment of a national curriculum and the focus on standardized tests and rankings to determine and publicly present a school's teaching quality have led to a massive narrowing of the education provided by schools in England (Alexander, 2010). A critical review by Stephen Gorard (2010), after thirty years of school effectiveness research, comes to the sobering conclusion that there is no reliable evidence of school effectiveness on which to base a serious school policy (see also Gorard, 2018; Hobbs, 2016; Strand, 2016).

The reform strategy of comprehensive bureaucratic-professional monitoring of schools in Germany

The publication of the first PISA study in 2000 caused a shock in Germany (cf. Baumert et al., 2001; Bieber et al., 2014; Ringarp, 2016; Waldow, 2019). The Standing Conference of the Ministers of Education and Cultural Affairs (KMK) of the German federal states (Länder) reacted to the PISA shock by extensively fixing educational standards and setting up a comprehensive battery of nationwide tests that go far beyond PISA competencies (Tillmann et al., 2008; Prenzel & Baumert, 2008; Prenzel et al., 2013, Reiss et al., 2016). A comprehensive system of educational monitoring has been established (Avenarius et al., 2003; Grünkorn et al., 2019). A commission of educational researchers provided the template for this program (Klieme et al., 2003, 2007, 2010). The "Ploen resolutions" of the KMK of June 2006 implemented a comprehensive system of "quality assurance" in the education system. The Institute for Quality Development in Education (IQB) in Berlin has taken a coordinating role in this process. Quality assurance requires the development of educational standards (Hartong, 2014). Such standards have been developed for the subjects of German and mathematics for the completion of primary school (grade 4), lower secondary school (grade 9), intermediate school leaving certificate (grade 10) and general higher education entrance qualification (grade 12/13); for the subjects of English and French for the completion of lower secondary school, intermediate school leaving certificate and general higher education entrance qualification; and for the subjects of biology, chemistry and physics for the intermediate school leaving certificate (Kultusministerkonferenz, 2015). In addition to the international tests, there is a whole series of national tests based on the educational standards that have been developed. The IQB conducts comparative national tests in German and mathematics every five years for primary school in grade 4, and every three years for lower secondary school, alternating between the language subjects (German, English, French) and the mathematical and scientific subjects (mathematics, biology, chemistry, physics). In addition, regular tests in these subjects are held in the individual federal states in different grades. It is important that the tests in the PISA subjects also go far beyond basic skills. In the languages, the focus is not only on reading comprehension, but also on free writing and listening, and in German also on language consideration and spelling (Kultusministerkonferenz, 2015).

These monitoring measures are common to all 16 states of the Federal Republic, which are, however autonomous in their educational policy. Nevertheless, there has always been coordination between the states, particularly based on the Standing Conference of the Ministers of Education and Cultural Affairs (KMK). This holds also for the implementation of reforms following GERM. There are differences between the 16 federal states in the strict implementation of measures ensuring increased monitoring and accountability. However, there is no "outlier", all 16 states share the policy of enhanced attention to such measures. This is what we scrutinize when looking at GERM implementation in Germany.

The next step was to change the educational process from a focus on educational content and qualifications to the acquisition of competences (Köller, 2008; Fleischer et al., 2013; Hartmann et al., 2017). As scientific preparation for this program, the German Research Foundation (DFG) established a research priority on competency models in 2006 (Klieme & Leutner, 2006; Leutner et al., 2017). The conversion of knowledge and qualifications into competences has, however, given rise to fierce criticism which, nevertheless, has not led to a departure from the path taken (Klein, 2010; Gruschka, 2013; Sander, 2013).

The practice of school policy at the level of the federal states proves that the old input-oriented bureaucratic control model has not been completely replaced but complemented by the new output-oriented model to form a hybrid. The granting of greater autonomy to schools and growing documentation requirements go hand in hand (Hartong, 2012, pp. 232–320; Rürup, 2019). The recommendations of the Standing Conference of Ministers of Education and Cultural Affairs (KMK) on

educational standards were implemented in Lower Saxony, for example, in such a way that in the subject of German the definition of standards based on the PISA test and the focus on reading skills that these standards are supposed to meet (Kultusministerkonferenz, 2013) were differentiated according to school types (Hauptschule, Realschule, Gymnasium) and referred to literary texts more than in PISA (MK Niedersachsen, 2006; Hartong, 2012, pp. 247–252). A key element of the changeover to output control was the introduction of independent schools in 2007 (Busemann, 2007). This includes the fact that each school has a school board, which is made up half of parents and half of pupils and forms a counterbalance to the teachers' conference. The principal of the independent school is the superior of the teachers and is responsible for the school's performance in quality controlled by the school inspection (Gatermann et al., 2010; Hartong, 2012, pp. 288–318).

What is the evidence regarding these reforms and regarding hypotheses H1 and H2? To begin with, the State Audit Office of Lower Saxony (Niedersächsischer Landesrechnungshof, 2016) made a scathing verdict on the "Independent School" project in its 2016 annual report and firmly called on the state government to correct this undesirable development. It is revealed that, contrary to the original intention, the project did not save costs, but rather generated exorbitant costs, 421 million Euros in ten years. And it is criticized that augmented school autonomy combined with tightened accountability places too much of a burden on teachers with documentation tasks and takes time away from teaching (Niedersächsischer Landesrechnungshof, 2016, p. 116).

The extent to which the reforms are associated with improving performance and decreasing performance gaps can be determined by comparing the performance of the ninth grade in mathematics and science in 2012 and 2018 using the IQB's Education Report of 2019. According to this report, deteriorations have occurred in several German federal states, both in the average achievement level and in the achievement gap between students with a lower or higher socioeconomic status or with or without a migrant background. There is a trend towards poorer performance among male pupils more than among female pupils. The lower performance level of pupils with a migrant background is largely explained by their lower socioeconomic status. Apparently, the German reform strategy of intensified educational monitoring has not been associated with enhancing performance and narrowing the performance gap between the socioeconomically better and worse off students. The Independent School in Lower Saxony has not achieved anything in the comparison between the federal states and between 2012 and 2018. This federal state is still slightly below the average of all federal states, and it deteriorated in several indicators from 2012 to 2018. Paradoxically, the central authority for comprehensive education monitoring has to conclude here that comprehensive education monitoring has not yet had the expected positive association with student performance (Stanat et al., 2019a, 2019b).

Is intensified monitoring of schools associated with decreasing socioeconomic achievement gaps? A multilevel regression analysis with data from PISA 2000, 2009 and 2015

As the literature reviewed so far proves, implementing elements of GERM on the specific developmental paths in the UK and in Germany is no success story. The available evidence speaks much more for hypothesis H2 than for hypothesis H1. A brief look at reform policies and PISA performance tells us that the UK and Germany are not unique cases in this respect. It is the overwhelming reality in the countries that have participated in increasing numbers since the first wave of 2000. Generally, the literature finds no closing of achievement gaps in 50 years (cf. Chmielewski, 2019; Early et al., 2019; Hanushek et al., 2019; Volante et al., 2019). There is not any single country which could serve as a real success story of GERM if we take PISA performance as yardstick. The PISA report of 2012 shows indeed a trend of increasing use of student assessment data for regional or national benchmarking and for monitoring teachers according to GERM in nearly all 38 respectively 37 countries for which data were available in 2003 and 2012 (OECD, 2013, pp. 160–161). However, if we look at the performance of countries in PISA, there is no significant improvement, particularly in high-income countries. This holds for the average PISA score, for the gap between the top and bottom socioeconomic quarter of students and the gap between the 5% highest and 5% lowest performers. Looking at countries belonging to the family of liberal welfare regimes which took a pioneering role in the implementation of GERM - Australia, New Zealand, Ireland, UK, U.S. – we see no improvement, but mostly deterioration. It may be that GERM policies help somewhat to establish an ordered form of teaching at all with regular class attendance in low-income countries, but they are not associated with any improvement in high-income countries, as also a recent OECD working paper confirms (Torres, 2021). More details are presented in online appendix F.

The country studies reviewed in the previous two sections and the brief look at the PISA performance of a larger number of countries inform us about reforms undertaken and about missing success. What they do not tell us directly is how much the governance tools as recommended by GERM and represented in hypothesis H1 and its opposite H2 do make a difference when we scrutinize the school and individual levels. Looking for an answer to this question is the aim of the following analysis. In this respect we aim at closing a research gap still not sufficiently eliminated by the existing literature.

Data and methods

We make use of the official PISA 2000, 2009 and 2015 dataset provided by the OECD (2020) and use R to calculate our statistical models.[1] Multilevel regression models are conducted to test the two hypotheses. The imputed PISA mean score of students in reading, mathematics and science serves as dependent variable.[2] To find out the differences between the countries at the three points in time, cross-sectional models are calculated for the United Kingdom and Germany for 2000, 2009 and 2015. Scotland is excluded in all three years as it was recorded separately in 2009. Furthermore, PISA scores could not be calculated for Wales in 2000 because no values were obtained for Welsh students in that wave. Speaking of the UK, therefore always means UK without Scotland in all three waves, and without Wales in 2000 only in the following. For securing reliability, we apply control and standardization measures of our data (cf. Jerrim et al., 2017).

Our multilevel regression models comprise of two levels. The first features variables on student level, the second on school level. We apply Snijders and Bosker's (1994) R^2 and the Intraclass Correlation Coefficient (ICC) (Bartko, 1966) to calculate the goodness of fit of our models. Snijders and Bosker's R^2 decomposes the variance explained by each model into variance explained on student level (level 1) and school level (level 2). As we apply the multiple imputation framework for missing data on the dependent variable, as well as on independent variables on student and school level, we report the combined R^2-values of both levels as well and use this combined measure to compare models. The ICC ranges from zero to one. The higher the ICC, the more similar to each other are the students at each school under observation regarding their imputed PISA mean scores. Furthermore, higher values prove the adequacy of making use of multilevel mixed effects models for our purpose. As we impute the models, we are also able to calculate significance values for our R^2 and ICC values, meaning that we are able to calculate whether the values differ from zero, therefore having additional proof

[1] We use the EdSurvey, tidyverse, dplyr, and haven packages for data preparation, psych, and BIFIEsurvey to conduct descriptive analysis, multilevel regression analysis, and postestimations.

[2] As the mean score as well as scores in reading literacy, mathematics and science are strongly correlated, it does not make a significant difference which score is taken.

for the adequacy of our models.

We apply z-standardization to improve the comparability of effect sizes. To compensate for missing observations on student level, we calculate student weights provided by the OECD. Shapiro-Francia W' tests were conducted to account for the normality of the error terms assumption. Additionally, we calculate the variance inflation factor (VIF) to test for heteroscedasticity (see appendix part D).

The models are constructed stepwise. We start with a *null model,* including the student- and school-specific intercepts of the imputed Pisa mean score only. Models one to three investigate measures of social inequality in both countries. *Model one* addresses the students' socioeconomic family background. We decided to use the index of economic, social and cultural status (ESCS) as provided by the PISA database to measure the impact of student's socioeconomic family background on imputed Pisa mean scores. The ESCS is defined as "composite score built by indicators of parental education, highest parental occupation, and home possessions including books" (OECD, 2017, pp. 339–340). For reasons of reliability, we decided to provide robustness checks using father's occupational status according to the international socioeconomic index of occupational status (ISEI) as indicator of a student's socioeconomic background (cf. Avvisati, 2020; Jerrim & Micklewright, 2014; Rutkowski & Rutkowski, 2013). For a detailed account, see appendix section E. The *second model* indicates social inequality between schools taking the average of the students' socioeconomic background per school as measure. The *third model* entails both measures of social inequality on student and school level and introduces an interaction between the student and school level measure of social inequality. By doing so, we simultaneously control if the school level effect is robust even if social inequality on student level is present. The *fourth model* investigates the association of the students' migration background and the PISA mean score. This model includes the students' migration status (second-generation versus no migration background), share of students with migration background at the respective school and whether the language spoken at home is the same as in the PISA test ('yes' as reference). The *fifth model* looks at the association between school climate and PISA mean scores. It includes variables on the disciplinary climate on the student and school levels. However, the operationalization of disciplinary climate varied between the three waves considered in this article. In 2000, the disciplinary climate during test language lessons was measured. In 2009, it was measured in general, whereas in 2015, data on the disciplinary climate in science classes was collected. Disciplinary climate is a dimensional variable constructed from items that measure whether students listen to the teacher or not, noisiness and disorderliness in class, the time it takes the students to be quiet and to participate in the lessons and the students' working conditions (OECD, 2017, p. 314). The *sixth model* is focused on variables of school governance: indices of (1) school autonomy, (2) school leadership and (3) school accountability according to hypotheses H1 and H2. See online appendix A for their construction. The *seventh model* entails variables regarding social inequality, migration background and school disciplinary climate. The*eighth model* further tests the robustness of variables indicating student and school ESCS as their interaction and governance variables are included. The *ninth model* scrutinizes the robustness of all variables hypothetically associated with PISA mean scores. The results are depicted in Tables 1 and 2. Descriptive statistics is provided in the online appendix, section B, for regression tables see section C, and for regression diagnostics see section D.

Results

Starting with the null model, we see high levels of homogeneity regarding the PISA mean scores in Germany (ICC = 0.630 in 2000, 0.577 in 2009, and 0.528 in 2015, $p < 0.001$), but lower levels in the UK (ICC = 0.288 in 2000, 0.275 in 2009, and 0.252 in 2015, $p < 0.001$). In both cases, these values drop sharply when social inequality among schools is measured, whereas other variables show a much lower impact on the ICC. The decline in the ICC tells that much of the homogeneity in PISA scores can be attributed to social inequality manifested at school level. However, we cannot draw inferences on the association of social inequality, migration background, school climate and school governance with the PISA mean score without investigating the models separately and taking the R^2 values into account.

Beginning with the association of student level ESCS with the PISA mean score in both Germany and the UK in model 1, we see positive, highly significant ($p < 0.001$) and robust effects across all models in 2000, 2009 and 2015. In comparison, individual student's ESCS is more strongly associated with the PISA score in the UK than in Germany. In the 2000 models, individual student's ESCS has an effect coefficient of β = 15.164 and explains 2.3% variance in Germany, whereas it explains 13.8% variance in the UK and shows a higher coefficient in model 1 (β = 32.124). The β value indicates that for each additional standard deviation of student level ESCS, the mean imputed Pisa score raises by 32.124 points in the UK and by 15.164 in Germany. In 2009, the coefficients and explained variance are β = 15.898 and 2.8% in Germany, β = 30.092 and

Table 1
United Kingdom, results of multilevel regression.

United Kingdom	2000		2009		2015	
	β when introduced	β full model	β when introduced	β full model	β when introduced	β full model
Student level: ESCS	32.124***	27.763***	30.092***	26.365***	23.601***	20.989***
School level: Mean ESCS	92.782***	62.924***	100.088***	73.35***	85.713***	64.553***
Student level migrant background 2nd generation	-3.371	-1.078	11.411	9.532	-2.801	0.008
School level % migrant background 2nd generation	52.454	30.619*	-59.546	-20.853	11.261	-4.428
Language at home = test language	34.735***	19.242**	24.382***	20.216***	15.016*	11.4
Student level school disciplinary climate	-2.34***	-1.983***	15,608***	14.251***	15.940***	14.992***
School level mean disciplinary climate	-1.529	0.776	19.839*	-17,268**	18.843	-3.993
School autonomy	15.903**	1.475	2.223	0.55	70.172	-23.215
Educational leadership	-767.067*	-351.271*	-47.352	-6.611	-4.356	0.334
Accountability	-11.311	-6.252*	-3.545	-1.134	-7.749**	-1.257
ICC	0.288***	0.087***	0.275***	0.092***	0.252***	0.081***
R^2 ESCS student level model	0.138***		0.136***		0.109***	
R^2 ESCS school level model	0.22***		0.212***		0.195***	
R^2 migrant background model	0.014		0.019		0.005	
R^2 disciplinary climate model	0.025		0.082***		0.091***	
R^2 governance tools model	0.026		0.011		0.06	
R^2 full model	0.315***		0.333***		0.312***	
Hypothesis H1	refuted		refuted		refuted	
Hypothesis H2	confirmed		confirmed		confirmed	

Explanation: Standardized beta coefficients $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; variables: first column shows beta coefficient of models focused on variable group, second column shows beta coefficient of full model.

Table 2
Germany, results of multilevel regression.

Germany	2000		2009		2015	
	β when introduced	β full model	β when introduced	β full model	β when introduced	β full model
Student level: ESCS	15.164***	10.343***	15.898***	10.23***	15.556***	10.589***
School level: Mean ESCS	135.492***	105.524***	130.167***	116.308***	110.349***	88.976***
Student level migrant background 2nd generation	-8.63*	-6.626*	-14.502***	-9.081*	-15.348**	-10.909
School level % migrant background 2nd generation	-3.396	-26.143	-167.545***	31.761	-114.044	-29.144
Language at home = test language	56.917***	47.739***	25.998***	22.666***	41.577***	39.731***
Student level school disciplinary climate	-1.001***	-0.946***	3.639**	2.913**	8.835***	8.018***
School level mean disciplinary climate	-1.031**	0.019	7.361	-13.821	63.45**	11.428
School autonomy	-3.375	-6.111	-3.559	-1.599	42.119	2.229
Educational leadership	-7.064	-178.001	115.195*	37.146	1.238	1.269
Accountability	-14.488*	-1.091	-8.358	-1.191	-3.693	-1.5
ICC	0.63***	0.25***	0.577***	0.265***	0.528***	0.18***
R^2 ESCS student level model	0.023***		0.028***		0.028**	
R^2 ESCS school level model	0.445***		0.44***		0.367***	
R^2 migrant background model	0.026*		0.098**		0.072	
R^2 disciplinary climate model	0.094***		0.004		0.074	
R^2 governance tools model	0.021		0.059		0.011	
R^2 full model	0.480***		0.475***		0.428***	
Hypothesis H1	refuted		refuted		refuted	
Hypothesis H2	confirmed		confirmed		confirmed	

Explanation: Standardized beta coefficients $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; variables: first column shows beta coefficient of models focused on variable group, second column shows beta coefficient of full model.

13.6% in the UK; in 2015, β =15,556 and 2.8% in Germany, β = 23.601 and 10.9% in the UK. These effects lose some of their strength when controlled in models 3, 7, and 9 in both countries in all three waves.

In contrast, social inequality between schools is a much more powerful predictor of the PISA mean score compared to student level ESCS in both countries in 2000, 2009 and 2015. This is surprising for the UK because this country has a comprehensive, competitive school system. It is not so for Germany because the German system is stratified into different ability tracks. In Germany in 2000, the effect coefficient of socioeconomic background on school level is nearly nine times stronger than the student level ESCS alone (β = 135.492, $p < 0.001$); it is robust when controlled and explains 44.5% variance present in the data in model 2. The effect is approximately the same in 2009 (β = 130.167, $p < 0.001$, explained variance 44%). These effects remain largely stable when controlled and are disproportionately high compared to all other variables included in our calculations. In 2015, the effect is only slightly weaker and also robust (β = 110.349, explained variance 36.7%). In the UK in 2000, model 2 reveals strong associations between school level ESCS and the PISA mean score. Albeit weaker than in Germany, social inequality between schools accounts for 22% variance in the data. Consequently, the effect coefficient is higher than on student level (β = 92.782, $p < 0.001$). Though a bit weaker, the effect is robust throughout all models. In 2009, both effect size and explained variance are similar and the effect is robust throughout all models (β = 100.088***, p <0.001, explained variance: 21.2%). In 2015, the effect is slightly weaker but still robust (β = 85.713, explained variance: 19.5%). These findings are mirrored in model 3 throughout all analyses. When both measures of social inequality as well as the interaction between both are included, the effect of social inequality remains stable. In fact, there are no interactions between student and school level ESCS present in Germany and the UK. This means that both levels contribute independently to the differences in Pisa scores.

Looking at the association between migration background and PISA mean scores in model 4, we see differences between Germany and the UK in explanatory power and partly in the direction of effects. Beginning with Germany in 2000, our model reveals negative, weakly significant associations between migration background on student level, and highly significant associations between language spoken at home and the PISA mean score. Having a migration background lowers the imputed mean Pisa score by 8.63 points ($p < 0.05$), whereas the test language spoken at home raises the score by 56.917 points ($p < 0.001$). In total, model 4 explains 2.6% variance ($p < 0.05$). Both indicators remain robust throughout models 7 and 9. This reveals a more direct relationship between migration background and test language spoken at home, on the one hand, and the PISA mean score, on the other hand. A slightly different pattern emerges in 2009. Here, model 4 explains 9.8% variance present in the data, and all three variables start as being highly significant. Individual migration background is negatively associated with the PISA mean score (β = -14.502, $p < 0.001$), school migration background even much more strongly (β = -167.545, $p < 0.001$), whereas test language spoken at home is positively related to the mean imputed Pisa score (β = 25.998, $p < 0.001$). Individual migration background is weakly robust, and school migration background is not robust when controlled for student and school level ESCS. Test language spoken at home remains highly significant and is robust throughout the models (β = 22.666, $p < 0.001$ in the full model). In 2015, there is no significant effect of school migration background on the PISA mean score, however a significant negative but not robust effect of student level migration background (β = -15.348, $p < 0.01$), and a significant positive and robust effect of test language spoken at home (β = 41.577, $p < 0.001$). In sum, migration status appears to be associated with lower socioeconomic status and therefore adds little to the explanation of the mean imputed PISA scores. Yet despite its interaction with socioeconomic status, the language spoken at home remains a viable predictor. In the UK, there is neither an effect of individual migration background nor an effect of school level migration background on the PISA mean score in all three waves. However, there is a positive, strongly significant and robust effect of test language spoken at home in 2000 and 2009 (β = 34.735, $p < 0.001$; β = 24.382, p< 0.001). In 2015, there is only a weak positive and not robust effect of test language spoken at home.

In comparison, we may attribute the differences in performance in Germany to social segregation and the immigration of socioeconomically deprived families (Strobel, 2016) with little competence in the German language and little knowledge about the German schooling system and its demands on the students (Kretschmer, 2019). In the United Kingdom, the lack of a negative effect of a migrant background may be due to the influx of families from the Commonwealth, who are linguistically and culturally close to the country of destination. And beyond the Commonwealth, English is spoken nearly everywhere. Therefore, students with migrant background perform hardly less well than native pupils in the PISA tests compared to immigrants in Germany (Hillmert, 2013).

In model 5, we see significant, but small and inconsistent effects of student level disciplinary climate on PISA mean scores in all three test

years in Germany. There is no effect on school level visible. We do not discuss this effect any further, as the variance explained does not differ significantly from zero. In the UK, there is a small negative, strongly significant and robust effect of individual disciplinary climate in 2000, which turns positive in 2009 and 2015. In 2000, the R^2 provided by model 5 does not differ significantly from zero, whereas in 2009 and 2015, the variance explained is 8.2% and 9.1% respectively. Additionally, school level disciplinary climate is weakly significant in 2009, but changes signs when controlled for student and school level ESCS.

Turning to the variables addressing the relationship between school governance and PISA mean scores according to hypothesis H1, that is indices of school autonomy, leadership, and accountability, in model 6, we observe mostly neither robust effects nor noteworthy portions of explained variance in our data in both countries. In fact, variance explained in both countries does not differ significantly from zero. Even if there might be partially significant effects, they do not account for any changes in the mean imputed PISA score. Overall, the results suggest that variables linked to school governance cannot compensate for effects of socioeconomic background. The differences between the UK and Germany may at least partly be due to the comprehensive school system in the UK and the stratified school system in Germany.

When controlled throughout models 7 to 9, the average mean ESCS on school level is the most powerful explanatory variable in both countries, more so in Germany than in the UK. For Germany in 2000, model 7 adds only 1.9% to explained variance compared to model 3 (which includes both levels and the interaction term), the three non-significant governance variables in model 8 add only 0.6%, the variables in full model 9 add 2.4%, exclusively due to the migration and disciplinary variables. In 2009, model 7 adds 1.6%, model 8 adds 0.5%, and model 9 adds 2% explained variance compared to model 2. In 2015, migration background and disciplinary variables in model 7 add 4.7% explained variance, the educational leadership, school autonomy, and accountability variables in model 8 add 0.5% explained variance, and the full model 4.8%.

In the UK, the socioeconomic status (student level plus school level plus interaction between both) in model 3 explains 29.2% of variance in the mean imputed PISA scores in 2000. Migration and disciplinary variables add 1.9% in model 7, whereas school governance variables add only 0.6% in model 8. In full model 9, all variables together add 2.3%. The figures in 2009 are 28.7% in model 3, 4.5% variance added in model 7, 0.1% in model 8, and 4.6% in the full model. In 2015, variance explained in model 3, is 25.2%, whereas models 7, 8 and 9 add 5.6%, 0.3%, and 6 % respectively. That means, student and school socioeconomic status together explain the highest portion of variance in the PISA mean score in both countries, migration and disciplinary variables add only a very small portion, governance variables add nothing at all.

These findings reveal a strong association between socioeconomic background, school segregation and school performance homogeneity in both countries, but more so in Germany as additionally reflected in the much higher ICC score in Germany at all three points in time (cf. Jenkins et al., 2008). This feature is matched by the achievement segregation associated with migrant background in Germany. Summing up, hypothesis H1 is refuted, and hypothesis H2 is confirmed.

Discussion and conclusion

In the United Kingdom, extensive reforms have been carried out from the 1980s onwards, the most extensive of which in England involved free choice of school in education markets, greater autonomy for schools and, in return, increased performance monitoring. The originally stratified school system has been dismantled to the point of maintaining only 163 grammar schools in favor of a comprehensive school system. As Scotland is not included in our PISA data, Scotland's resistance to follow England's GERM reforms fully does not affect our analysis and interpretation of the data. The empirical studies we have consulted prove that the forced promotion of free choice of school in education markets is not associated with increasing educational achievement but rather with either reproduced or even augmented existing educational inequalities. The greater autonomy of schools, coupled with tightened performance monitoring, has massively increased the documentation work of school administrators and teachers and narrowed down teaching to central subjects and test preparation. In Germany, school monitoring has been expanded, and in some federal states, such as Lower Saxony, the autonomy of schools has been strengthened while at the same time tightening self-control and accountability. However, education markets, such as those in the United Kingdom, have not been established, and the stratified school system with the Gymnasium at the top has been maintained. In contrast to the UK, there are fewer empirical studies on the outcomes of the reform measures. However, the IQB's education reporting clearly proves that there have been no significant improvements in performance and that there are still large gaps in achievement based on family background.

According to the PISA data (OECD, 2020), no association of the reforms undertaken can be identified, neither of the market-based nor of the bureaucratic-professional reforms aimed at extended monitoring. This is true both in the longitudinal look at the test years since 2000 and in the comparison of the two countries with each other and with competing countries. The longitudinal view shows a high level of continuity in PISA test performance, limited only by individual outliers. And these outliers cannot be attributed to specific reform measures. In comparison with other countries, both the United Kingdom and Germany have specific distinguishing features, which, however, have no discernible connection to the country-specific reform measures. Rather, country-specific structural and cultural characteristics are noticeable, which are sedimented in the institutions of both school systems (cf. Meyer & Schiller, 2013; Tienken et al., 2017). Socioeconomic family background overwhelmingly explains PISA performance throughout all scrutinized years, and GERM policies such as increased school autonomy accompanied by strong leadership and strict accountability do not remove this association (see online appendix, section F). As the coverage of youth 15 years old in the PISA tests is lower in the UK than in Germany, average performance might be overestimated and the association of socioeconomic status with PISA scores underestimated in the case of the UK (OECD, 2010a: Tab. A2.1; 2016a: Tab. A2.1; 2001: Tab. A3.1; cf. Anders et al., 2019). Jerrim (2021), (p. 18) concludes for PISA 2018 that only 61% of eligible students participated in the test in the UK as against 88% in Germany. The UK is at the bottom regarding participation rate in the test. Jerrim (2021), (p. 3, 25) provides hints to biased participation insofar as lower-performing pupils were underrepresented in the final sample and speaks of an overestimation of 10 to 15 points for England and Wales. With also lower coverage rates reported for the UK compared to Germany in the test waves scrutinized in our study, there might indeed be an overestimation of UK performance in our results.

Comparing the four jurisdictions of the UK, England, Scotland and Northern Ireland are around, slightly above or slightly below the OECD average of around 490 points. Wales is below OECD average (cf. Jerrim, 2021, p. 2). In Germany, there are differences in average performance and achievement gaps between the 16 states of the Federal Republic. However, they are not reported publicly for PISA. The monitoring of the IQB clearly shows such differences. There are good reasons for assuming that the differences would not be different with PISA. The IQB report shows predominantly differences which are associated with wealth, class structure and strength of the school system in terms of lower or higher standards and lower or higher enforced discipline (cf. Stanat et al., 2019a, 2019b).

Our multilevel linear regression analysis of PISA data has not generated any evidence that the market-based or bureaucratic-professional reform strategies aimed at comprehensive monitoring of schools are associated with visible achievements in the UK or Germany. The socioeconomic status of the school attended and of the student as well as to a much smaller extent school discipline are of crucial importance. This result is in line with research showing that a school's

socioeconomic status is the primary factor associated with the performance level of students, varying according to structural and institutional features of school systems such as tracking age and share of public schooling in the system (Perry & McConney, 2013; Perry & Lubienski, 2020; Sciffer et al., 2021). The expansion of secondary and tertiary education having been completed in the 1990s, the influence of family background on educational achievement has gained in importance again. In Germany, this is still evident in a school system stratified into ability tracks (Stocké et al., 2019), but now also in a growing achievement differentiation in the Gymnasium. In the comprehensive school system of the UK, it is evident in the achievement differentiation within schools and between schools according to the average socioeconomic status of their students. However, we do not see an extension of the association of the socioeconomic background of students with their educational achievement from 2000 to 2015. In our study, the association of these factors with student performance is stable to slightly decreased in Germany and the UK in comparison of 2000 and 2015. The changes are not large enough to speak of a systematically reduced association of socioeconomic factors with student performance, particularly in face of possibly biased participation in PISA in the UK which might imply underestimation of achievement gaps. And basically, governance variables are not strong enough in their association with performance that we can think of their contribution to these little changes.

The extent to which our results are robust must be demonstrated by further studies that use data other than PISA data, look at other countries and further PISA test years and include other structural, cultural, and governance-specific factors in the analysis. There is the question whether PISA results are distorted by varying student motivation in the test. The OECD report on PISA 2018 entails an annex dealing with this question on the country level (OECD, 2019a, Annex A8). According to this report, in the OECD average, 68% of the students surveyed estimated their effort in the test to be somewhat lower compared to a test counting to their mark at school, that is at 7.6 instead of about 9 on a scale ranging from 0 to 10. In the UK, the percentage and the scale value were 75 and 7.5, in Germany 80 and 7.2. There is not the same information on the PISA waves of 2000, 2009 and 2015, but we may assume that the picture would be rather the same. With a lower motivation of one point on a ten-point scale we may guess that PISA scores of countries would be up to 10 percent higher if PISA were a high-stakes test and more so in countries with higher difference between low-stakes and high-stakes testing. According to the 2018 survey, this would reduce the distance between the East Asian top performers and the Western countries ranging around the OECD average somewhat. As the difference between Germany and the UK is small in this respect, there should be no effect on our comparison between these two countries. However, there are good reasons to take socioeconomic status and discipline on student and school level as proxies for student motivation. PISA reports on student engagement in learning give some hints in this direction (see, e. g., OECD 2003, p. 296; 2010b, pp. 138–141; 2016a, pp. 338–339). That means, on the school and student levels, student motivation is indirectly taken into account in our regression analysis. There are doubts about the quality of this data due to the tendency of surveyed school principals to simplify their answering behavior (Blasius & Thiessen, 2015). And PISA results need careful interpretation to avoid fallacies (Araujo et al., 2017; Morgan, 2017; Rutkowski & Rutkowski, 2016). We decided to use a multiple imputation framework in order to cope with the sampling problems in PISA estimations and the potential biases in the estimators. On the one hand, this technique can be used to fill in missing values of students based on shared characteristics. On the other hand, the weights provided by the OECD form a corrective for the calculated standard errors and variance elucidation of multilevel models. It even allows us to test if our goodness of fit measures (R^2- and ICC-values) are significantly different from zero and thus the variables included in our models can be used to explain differences in PISA outcomes for both the UK and Germany. For these reasons, the data should be adequate for answering our research questions.

For school policy, our results suggest that the global reform agenda, which relies on increasing school autonomy along with tightened educational monitoring should be examined more critically regarding promised achievements and undesired side effects. As international benchmarking advanced by PISA as major agent of GERM has not really improved quality and equity in education, critical assessments of PISA are all the more important to be recognized.

Funding

Funding for research published in this article was provided by the Deutsche Forschungsgemeinschaft under project number 429041218.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijedro.2022.100165.

References

Alexander, R. (Ed.). (2010). *Children, their world, their education: Final report and recommendations of the Cambridge primary review*. New York and London: Routledge.
Allen, R., & Higham, R. (2018). Quasi-markets, school diversity and social selection: Analysing the case of free schools in England, five years on. *London Review of Education, 16*(2), 191–213.
Anders, J., Has, S., Jerrim, J., Shure, N., & Zieger, L. (2019). *Is Canada really an education superpower? The impact of exclusions and non-response on results from PISA 2015*. UCL Institute of Education. *Working Paper*, No. 19-11.
Angus, L. (2015). School choice: Neoliberal education policy and imagined futures. *British Journal of Sociology of Education, 36*(3), 395–413.
Araujo, L., Saltelli, A., & Schnepf, S. V. (2017). Do PISA data justify PISA-based education policy? *International Journal of Comparative Education and Development, 19*(1), 20–34.
Avenarius, H., Ditton, H., Döbert, H., Klemm, K., Klieme, E., Rürup, M., & Weiß, M. (2003). *Bildungsbericht für Deutschland: Erste Befunde*. Opladen: Leske + Budrich.
Avvisati, F. (2020). The measure of socioeconomic status in PISA: A review and some suggested improvements. *Large Scale Assessments in Education, 8*(8), 1–37.
Ball, S. J. (1993). Education markets, choice and social class: The market as a class strategy in the UK and the USA. *British Journal of Sociology of Education, 14*(1), 3–19.
Ball, S. J. (2008). *The education debate*. Bristol: The Policy Press.
Ball, S. J. (2012). *Global education Inc. New policy networks and the neoliberal imaginary*. London and New York: Routledge.
Ball, S. J., Bowe, R., & Gewirtz, S. (1996). School choice, social class and distinction: The realization of social advantage in education. *Journal of Education Policy, 11*(1), 89–112.
Ball, S. J., Maguire, M., Braun, A., Perryman, J., & Hoskins, K. (2012). Assessment technologies in schools: Deliverology' and the play of dominations. *Research Papers in Education, 27*(5), 513–533.
Barber, M., & Mourshed, M. (2007). *How the worlds' best-performing school systems come out on top*. London: McKinsey & Company.
Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19*(1), 3–11.
Baumert, J., et al. (Eds.). (2001). *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.
Bieber, T., Martens, K., Niemann, D., & Windzio, M. (2014). Grenzenlose Bildungspolitik? Empirische Evidenz für PISA als weltweites Leitbild für nationale Bildungsreformen. *Zeitschrift für Erziehungswissenschaft, 17*(4), 141–166.
Blasius, J., & Thiessen, V. (2015). Should we trust survey data? Assessing response simplification and data fabrication. *Social Science Research, 52*, 479–493.
Boudon, R. (1974). *Education, opportunity, and social inequality. Changing prospects in western society*. New York: Wiley.
Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste*. Cambridge, MA: Harvard University Press.
Bourdieu, P. (1998). *The state nobility: Elite schools in the field of power*. Stanford, CA: Stanford University Press.
Bourdieu, P., & Passeron, J. C. (1990). *Reproduction in education, society and culture*. London: Sage.
Breen, R., & Jonsson, J. O. (2005). Inequality of opportunity in comparative perspective: Recent research on educational attainment and social mobility. *Annual Review of Sociology, 31*(1), 223–243.
Burgess, S., Greaves, E., & Vignoles, A. (2019). School choice in England: Evidence from national administrative data. *Oxford Review of Education, 45*(5), 690–710.
Busemann, B. (2007). Eigenverantwortliche Schule: Der Weg aus dem PISA-Tal. In B. Busemann, J. Oelkers, & H. S. Rosenbusch (Eds.), *Eigenverantwortliche Schule – Ein Leitfaden. Konzepte, Wege, Akteure* (pp. 3–10). Köln: Luchterhand.
Chmielewski, A. K. (2019). The global increase in the socioeconomic achievement gap, 1964 to 2015. *American Sociological Review, 84*(3), 517–544.
Courtney, S. J. (2016). Post-panopticism and school inspection in England. *British Journal of Sociology of Education, 37*(4), 623–642.

Courtney, S. J., & Gunter, H. M. (2015). Get off my bus! School leaders, vision work and the elimination of teachers. *International Journal of Leadership in Education, 18*(4), 395–417.
Croxford, L., & Raffe, D. (2007). Education markets and social class inequality. In R. Teese, S. Lamb, M. Duru-Bellat, & S. Helme (Eds.), *International studies in educational inequality, theory and policy* (pp. 710–737). Dordrecht: Springer.
Early, E., Miller, S., Dunne, L., Thurston, A., & Filiz, M. (2019). The influence of socio-economic background and gender on PISA attainment in the United Kingdom: A systematic review. *Review of Education, 8*(1), 120–152.
Ellis, V., Steadman, S., & Trippestad, T. A. (2019). Teacher education and the GERM: Policy entrepreneurship, disruptive innovation and the rhetorics of reform. *Educational Review, 71*(1), 101–121.
Erikson, R., & Goldthorpe, J. H. (1992). Individual or family? Results from two approaches to class assignment. *Acta Sociologica, 35*(2), 95–105.
Eyles, A., Hupkau, C., & Machin, S. (2016). School reforms and pupil performance. *Labour Economics, 41*, 9–19.
Fleischer, J., Koeppen, K., Kenk, M., Klieme, E., & Leutner, D. (2013). Kompetenzmodellierung: Struktur, Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. *Zeitschrift für Erziehungswissenschaft, 16*(1), 5–22.
Gatermann, K., Gierke, J., Poppinga, A., & von Saldern, M. (2010). SEIS – Kritische Darstellung eines Selbstevaluationsinstruments. In M. von Saldern (Ed.), *Selbstevaluation von Schule. Hintergrund – Durchführung – Kritik* (pp. 101–126). Norderstedt: Books on Demand.
Goldthorpe, J. (2007a). 'Cultural capital.' Some critical observations. *Sociologica, 1*(2), 1–23.
Goldthorpe, J. (2007b). Outline of a theory of social mobility. In J. Goldthorpe (Ed.), *On sociology. Vol. I: Critique and program* (pp. 154–185). Stanford, CA: Stanford University Press.
Gorard, S. (2010). Serious doubts about school effectiveness. *British Educational Research Journal, 36*(5), 745–766.
Gorard, S. (2014). The link between academies in England, pupil outcomes and local patterns of socioeconomic segregation between schools. *Research Papers in Education, 29*(3), 268–284.
Gorard, S. (2018). *Education Policy, Equity And Effectiveness: Evidence Of Equity And Effectiveness*. Bristol: Policy Press.
Gorard, S. (2019). *School choice in an established market*. London and New York: Routledge.
Greany, T., & Higham, R. (2018). *Hierarchy, markets and networks: Analysing the 'self-improving school-led system' Agenda in England and the implications for schools*. London: UCL Institute of Education Press.
Greany, T., & Waterhouse, J. (2016). Rebels against the system: Leadership agency and curriculum innovation in the context of school autonomy and accountability in England. *International Journal of Educational Management, 30*(7), 1188–1206.
Grünkorn, J., Klieme, E., & Stanat, P. (2019). Bildungsmonitoring und Qualitätssicherung. In O. Köller (Ed.), *Das Bildungswesen in Deutschland. Bestand und Potenziale* (pp. 245–280). Bad Heilbrunn: Julius Klinkhardt.
Gruschka, A. (2013). Reformrückschlag. Über einen Versuch, Kompetenz zum kompetenzorientierten Unterricht zu vermitteln. *Pädagogische Korrespondenz, 48* (Fall), 58–78.
Gunter, H. M., & McGinity, R. (2014). The politics of the academies programme: Natality and pluralism in education policy-making. *Research Papers in Education, 29*(3), 300–314.
Hansen, K. Y., & Gustafsson, J. E. (2019). Identifying the key source of deteriorating educational equity in Sweden between 1998 and 2014. *International Journal of Educational Research, 93*, 79–90.
Hanushek, E. A., Link, S., & Woessmann, L. (2013). Does school autonomy make sense everywhere? Panel estimates from PISA. *Journal of Development Economics, 104*, 212–232.
Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessmann, L. (2019). The achievement gap fails to close: Half century of testing shows persistent divide between haves and have-nots. *Education Next, 19*(3), 8–18. S.
ed. by Hartmann, U., Decristan, J., & Klieme, E. (2017). Unterricht als Feld evidenzbasierter Bildungspraxis? In J. Baumert, & K. J. Tillmann (Eds.), *Empirische Bildungsforschung* (pp. 179–199). Wiesbaden: Springer VS. ed. by.
Hartong, S. (2012). *Basiskompetenzen statt Bildung? Wie PISA die deutschen Schulen verändert hat*. Frankfurt/New York: Campus.
Hartong, S. (2014). Neue Bildungsregulierung im Zeitalter der „governance by numbers": Das Beispiel standardisierter Bildungsreformen in Deutschland und den USA. *Leviathan, 42*(4), 606–634.
Heilbronn, R. (2016). Freedoms and perils: Academy schools in England. *Journal of Philosophy of Education, 50*(3), 306–318.
Hertel, F. R., & Groh-Samberg, O. (2019). The relation between inequality and intergenerational class mobility in 39 countries. *American Sociological Review, 84*(6), 1099–1133.
Hillmert, S. (2013). Links between immigration and social inequality in education: A comparison among five European countries. *Research in Social Stratification and Mobility, 32*(June 2013), 7–23.
Hobbs, G. (2016). Explaining social class inequalities in educational achievement in the UK: Quantifying the contribution of social class differences in school 'effectiveness'. *Oxford Review of Education, 42*(1), 16–35.
Jenkins, S. P., Micklewright, J., & Schnepf, S. V. (2008). Social segregation in secondary schools: How does England compare with other countries? *Oxford Review of Education, 34*(1), 21–37.
Jerrim, J. (2021). PISA 2018 in England, Northern Ireland, Scotland and Wales: Is the data really representative of all four corners of the UK? *Review of Education, 9*(3), e3270.
Jerrim, J., & Micklewright, J. (2014). Socioeconomic gradients in children's cognitive skills: Are cross-country comparisons robust to who reports family background? *European Sociological Review, 30*(6), 766–781.
Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., & Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review, 61*(December 2017), 51–58.
Klein, H. P. (2010). Die neue Kompetenzorientierung: Exzellenz oder Nivellierung? *Zeitschrift für Didaktik der Biowissenschaften, 1*(1), 15–26.
Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. *Zeitschrift für Pädagogik, 52*(6), 876–903.
Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., & Wolfgang, S. (2010). *PISA 2009: Bilanz nach einem Jahrzehnt*. Münster: Waxmann Verlag.
Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Vollmer, J., Tenorth, H. E., & Vollmer, J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Reihe Bildungsreform, 1*. Bonn: Bundesministerium für Bildung und Forschung.
Klieme, E., Avenarius, H., Baethge, M., Döbert, H., Hetmeier, H. W., Meister-Scheufelen, G., Rauschenbach, T., & Wolter, A (2007). Grundkonzeption der Bildungsberichterstattung für Deutschland. In H. H. Krüger, T. Rauschenbach, & U. Sander (Eds.), *Bildungs- und Sozialberichterstattung* (pp. 129–145). Wiesbaden: VS Verlag.
Kretschmer, D. (2019). Explaining native-migrant differences in parental knowledge about the German educational system. *International Migration, 57*(1), 281–297.
Kultusministerkonferenz. (2013). Bildungsstandards der Kultusministerkonferenz. www.kmk.org (accessed 28 October 2016).
Kultusministerkonferenz. (2015). Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf (accessed 11 April 2020).
Köller, O. (2008). Bildungsstandards in einem Gesamtsystem der Qualitätssicherung im allgemeinbildenden Schulsystem Deutschlands. In Klieme, E., & Tippelt, R. (Eds.). Qualitätssicherung im bildungswesen, Zeitschrift für Pädagogik, 54, 53. Beiheft (pp. 59–75). Weinheim, Basel: Beltz.
Leutner, D., Fleischer, J., Grünkorn, J., & Klieme, E. (2017). *Competence assessment in education. Research, models and instruments*. Hamburg: Springer.
Lubienski, C., Perry, L. B., Kim, J., & Canbolat, Y. (2021). Market models and segregation: Examining mechanisms of student sorting. *Comparative Education*. https://doi.org/10.1080/03050068.2021.2013043
Meyer, H. D., & Schiller, K (2013). Gauging the role of non-educational effects in large-scale assessments: Socioeconomics, culture and PISA outcomes. In H. D. Meyer, & A. Benavot (Eds.), *PISA, power, and policy: The emergence of global educational governance* (pp. 207–224). Oxford: Symposium Books.
Münch, R. (2020). *Governing the school under three decades of neoliberal reform. From educracy to the education-industrial complex*. London and New York: Routledge.
Niedersachsen, MK (2006). *Orientierungsrahmen Schulqualität in Niedersachsen*. Hannover. https://www.mk.niedersachsen.de/startseite/schule/schulqualitat/orientierungsrahmen_schulqualitat_in_niedersachsen/orientierungsrahmen—basis-fuer-schulqualitaet-in-niedersachsen-6339.html (accessed 15 March 2020).
Morgan, C. (2017). Problematizing the expansion of PISA's explanatory power. In L. Volante (Ed.), *The PISA effect on global educational governance* (pp. 17–31). London and New York: Routledge.
Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better*. London: McKinsey & Company.
Niedersächsischer Landesrechnungshof. (2016). *Jahresbericht des Niedersächsischen Landesrechnungshofs 2016*. Hannover.
OECD. (2001). *Knowledge and skills for Life. First results from the OECD programme for international student assessment (PISA) 2000*. Paris: OECD.
OECD. (2003). *Literacy skills for the world of tomorrow. Further results from PISA 2000*. Paris: OECD.
OECD. (2010a). *PISA 2009 results, I*. Paris: OECD.
OECD. (2010b). *PISA 2009 results, III*. Paris: OECD.
OECD. (2011). *Strong performers and successful reformers in education: Lessons from PISA for the United States*. Paris: OECD.
OECD. (2013). PISA 2012 results. In *What makes schools successful? Resources, policies and practices, IV*. Paris: OECD.
OECD. (2015). *Improving schools in Sweden: An OECD Perspective*. Paris: OECD.
OECD. (2016a). *PISA 2015 Results, I*. Paris: OECD.
OECD. (2016b). *PISA 2015 Results, II*. Paris: OECD.
OECD. (2016c). *PISA 2015 Results, III*. Paris: OECD.
OECD. (2017). *PISA 2015 technical report*. Paris: OECD.
OECD. (2019a). *PISA 2018 Results, I*. Paris: OECD.
OECD. (2019b). *PISA 2018 Results, III*. Paris: OECD.
OECD. (2020). *PISA 2000-2018 database*. Paris: OECD Publishing. https://www.oecd.org/pisa/data/2018database/ accessed 10 March 2020.
OPSI. (1989). *Education reform Act 1988*. London: Her Majesty's Stationery Office.
Perry, L. B., & McConney, A. (2013). School socioeconomic status and student outcomes in reading and mathematics: A comparison of Australia and Canada. *Australian Journal of Education, 57*(2), 124–140.
Perry, L., & Lubienski, C. (2020). Between-school stratification of academic curricular offerings in upper secondary education: School decision-making, curriculum policy context, and the educational marketplace. *Oxford Review of Education*, 1–19. https://doi.org/10.1080/03054985.2020.1739012
Perry, T. (2017). Inter-method reliability of school effectiveness measures: A comparison of value-added and regression discontinuity estimates. *School Effectiveness and School Improvement, 28*(1), 22–38.

Perryman, J. (2006). Panoptic performativity and school inspection regimes: Disciplinary mechanisms and life under special measures. *Journal of Education Policy, 21*(2), 147–161.

Perryman, J., Maguire, M, Braun, A., & Ball, S. (2018). Surveillance, governmentality and moving the goalposts: The influence of Ofsted on the work of schools in a post-panoptic era. *British Journal of Educational Studies, 66*(2), 145–163.

Prenzel, M., & Baumert, J. (Eds.). (2008). *Vertiefende Analysen zu PISA 2006. Sonderheft 10 der Zeitschrift für Erziehungswissenschaft*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Prenzel, M., Sälzer, C., Klieme, E., & Köller, O. (2013). *PISA 2012. Fortschritte und Herausforderungen in Deutschland*. Münster: Waxmann.

Raffe, D., Brannen, K., Croxford, L., & Martin, C. (1999). Comparing England, Scotland, Wales and Northern Ireland: The case for 'home internationals' in comparative research. *Comparative Education, 35*(1), 9–25.

Rayner, S. M., Courtney, S. J., & Gunter, H. M. (2018). Theorising systemic change: Learning from the academisation project in England. *Journal of Education Policy, 33*(1), 143–162.

Reiss, K., Sälzer, C., Schiepe-Tiska, A., Klieme, E., & Köller, O. (2016). *PISA 2015. Eine Studie zwischen Kontinuität und Innovation*. Münster: Waxmann.

Reynolds, D. (2010). *Failure-free education? The past, present and future of school effectiveness and school improvement*. London and New York: Routledge.

Reynolds, D., & Mckimm, J. (2020). Education in Wales: Attempting systemic reform to combat underachievement, 2000–2018. In A. Jones, & M. Harris (Eds.), *Leading and transforming education systems* (pp. 133–148). Singapore: Springer.

Ringarp, J. (2016). PISA lends legitimacy: A study of education policy changes in Germany and Sweden after 2000. *European Educational Research Journal, 15*(4), 447–461.

Rürup, M. (2019). Schulautonomie in Deutschland–ein Dauerthema der Schulreform. In E. Rauscher, C. Wiesner, D. Paasch, & P. Heißenberger (Eds.), *Schulautonomie–Perspektiven in Europa: Befunde aus dem EU-Projekt INNOVITAS* (pp. 61–75). Münster: Waxmann.

Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education, 8*(3), 259–278.

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher, 45*(4), 252–257.

Sahlberg, P. (2015). *Finnish lessons 2.0. What can the world learn from educational change in Finland?* (2nd ed.). New York: Teachers College Press.

Sahlberg, P. (2016). The global educational reform movement and its impact on schooling. In K. Mundy, A. Green, B. Lingard, & A. Verger (Eds.), *The handbook of global education policy* (pp. 128–144). Chichester: Wiley Blackwell.

Salokangas, M., & Ainscow, M. (2017). *Inside the autonomous school: Making sense of a global educational trend*. London and New York: Routledge.

Sander, W. (2013). Die Kompetenzblase – Transformationen und Grenzen der Kompetenzorientierung. *Zeitschrift für Didaktik der Gesellschaftswissenschaften, 4*(1), 100–124.

Sciffer, M. G., Perry, L. B., & McConney, A. (2021). Does school socioeconomic composition matter more in some countries than others, and if so, why? *Comparative Education*. https://doi.org/10.1080/03050068.2021.2013045

Sjøberg, S. (2019). The PISA-syndrome – How the OECD has hijacked the way we perceive pupils, schools and education. *Confero: Essays on Education, Philosophy and Politics, 7*(1), 12–65.

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research, 22*(3), 342–363.

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., & Henschel, S. (2019a). *Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I im zweiten Ländervergleich*. Münster: Waxmann.

Stanat, P., Schipolowski, S., Mahler, N., Weirich, S., & Henschel, S. (2019b). *IQB trends in student achievement 2018. The second national assessment of mathematics and science proficiency at the end of ninth grade*. Münster: Waxmann.

Stocké, V., Blossfeld, H. P., Hoenig, K., & Sixt, M (2019). Social inequality and educational decisions in the life course. In H. P. Blossfeld, & H. G. Rossbach (Eds.), *Education as a lifelong process: The German national educational panel study (NEPS)* (pp. 101–118). Wiesbaden: Springer.

Strand, S. (2016). Do some schools narrow the gap? Differential school effectiveness revisited. *Review of Education, 4*(2), 107–144.

Strobel, B. (2016). Does family language matter? The role of foreign language use and family social capital in the educational achievement of immigrant students in Germany. *Ethnic and Racial Studies, 39*(14), 2641–2663.

Tienken, C. H., Colella, A., Angelillo, C., Fox, M., McCahill, K. R., & Wolfe, A. (2017). Predicting middle level state standardized test results using family and community demographic data. *RMLE Online, 40*(1), 1–13.

Tillmann, K. J., Dedering, K., Kneuper, D., Kuhlmann, C., & Nessel, I. (2008). *PISA als bildungspolitisches Ereignis. Fallstudien in vier Bundesländern*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Pearson. (2012). The learning curve. Lessons in country performance in education. 2012 Report. https://issuu.com/pearsonmiddleeast/docs/learningcurve_report.www.thelearningcurve.pearson.com (accessed 7 September 2020).

Torres, R. (2021). Does test-based school accountability have an impact on student achievement and equity in education? A panel approach using PISA. OECD Education Working Paper No. 250. https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/WKP(2021)7&docLanguage=En (accessed 10 February 2022).

Van Zanten, A. (2005). New modes of reproducing social inequality in education: The changing role of parents, teachers, schools and educational policies. *European Educational Research Journal, 4*(3), 155–169.

Verger, A., Parcerisa, L., & Fontdevila, C. (2019). The growth and spread of large-scale assessments and test-based accountabilities: A political sociology of global education reforms. *Educational Review, 71*(1), 5–30.

Volante, L., Schnepf, S. V., Jerrim, J., & Klinger, D. A. (2019). *Socioeconomic inequality and student outcomes: Cross-national trends, policies, and practices*. Singapore: Springer.

Waldow, F (2019). Pedagogical paradise' and 'exam hell': PISA top scorers as projection screens in German print media. In F. Waldow, & G. Steiner-Kahmsi (Eds.), *Understanding PISA's attractiveness: Critical analyses in comparative policy studies* (pp. 65–88). London and New York: Bloomsbury.

Whitty, G., Power, S., & Halpin, D. (1998). *Devolution and Choice in Education. The School, the State, and the Market*. Buckingham: Open University Press.

Woods, P. A., Torrance, D., Donnelly, C., Hamilton, T., Jones, K., & Potter, I. (2021). Constructions and purposes of school leadership in the UK. *School Leadership & Management, 41*(1-2), 152–170.

You, Y. (2017). Comparing school accountability in England and its East Asian sources of 'borrowing'. *Comparative Education, 53*(2), 224–244.