

## 'What works' depends: teacher accountability policy and sociocultural context in international large-scale surveys

Hwa, Yue-Yi

Veröffentlichungsversion / Published Version

Zeitschriftenartikel / journal article

### Empfohlene Zitierung / Suggested Citation:

Hwa, Y.-Y. (2021). 'What works' depends: teacher accountability policy and sociocultural context in international large-scale surveys. *Journal of Education Policy*, 1-26. <https://doi.org/10.1080/02680939.2021.2009919>

### Nutzungsbedingungen:

Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:

<https://creativecommons.org/licenses/by/4.0/deed.de>

### Terms of use:

This document is made available under a CC BY Licence (Attribution). For more information see:

<https://creativecommons.org/licenses/by/4.0>



## 'What works' depends: teacher accountability policy and sociocultural context in international large-scale surveys

Yue-Yi Hwa

To cite this article: Yue-Yi Hwa (2021): 'What works' depends: teacher accountability policy and sociocultural context in international large-scale surveys, Journal of Education Policy, DOI: [10.1080/02680939.2021.2009919](https://doi.org/10.1080/02680939.2021.2009919)

To link to this article: <https://doi.org/10.1080/02680939.2021.2009919>



© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 20 Dec 2021.



[Submit your article to this journal](#)



Article views: 955



[View related articles](#)



[View Crossmark data](#)



Citing articles: 1 [View citing articles](#)

# 'What works' depends: teacher accountability policy and sociocultural context in international large-scale surveys

Yue-Yi Hwa 

Faculty of Education, University of Cambridge, Cambridge, UK

## ABSTRACT

Despite a growing emphasis in education policy on 'what works for whom and in what circumstances', there is still considerable attention to decontextualised 'best practices' that emerge from cross-country comparisons of student achievement. Also, while operational and even political aspects of context are increasingly incorporated into policy research, there is relatively little attention to the relationship between sociocultural context and education policy. In this paper, I explore the extent to which national sociocultural context influences the relationship between one aspect of policy – teacher accountability – and student outcomes. I do so by using multilevel modelling to analyse international survey data on education (from PISA 2012, PISA 2015, and TIMSS 2015) matched at the country level with survey data on culture (from the World Values Survey and Hofstede's IBM study). I find that one of the sociocultural constructs significantly and consistently moderates the relationship between teacher accountability and student outcomes, suggesting that some teacher accountability approaches may be beneficial in certain sociocultural contexts but detrimental in others. This finding implies a need for caution in generating universal policy prescriptions from international assessments such as PISA and TIMSS. It also strengthens the case for viewing teacher accountability as a socioculturally embedded process.

## ARTICLE HISTORY

Received 19 April 2021

Accepted 18 November 2021


## KEYWORDS

Teacher accountability;  
sociocultural context;  
international large-scale  
student assessments; PISA;  
education policy

## Introduction

In both the discourse and design of teacher accountability policy, there is frequently a tension between the utility of standardisation and the reality of context-specific variation across country, school, and classroom contexts. This tension is apparent in, on one hand, the growing popularity of externally regulated accountability approaches and, on the other, the increasingly widespread recognition that accountability policy implementation is strongly contingent on contexts, relationships, and perceptions (Verger and Parcerisa 2017; see also UNESCO 2017). Put differently, there is substantial attention to teacher accountability policy, alongside growing awareness that it is not just a matter of 'what works', but rather of 'what works for whom and in what circumstances' (Pawson and Tilley 1997, 144).

**CONTACT** Yue-Yi Hwa  [yue-yi.hwa@bsg.ox.ac.uk](mailto:yue-yi.hwa@bsg.ox.ac.uk)  Blavatnik School of Government, University of Oxford, Oxford, OX2 6GG, United Kingdom

 Supplemental data for this article can be accessed [here](#).

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This tension between the attractiveness of the general and the nuances of the specific also affects the domain of cross-country learning in education policy. The past few decades have seen increasing interest in, and perceived legitimacy from, reproducing the policy choices of celebrated education systems (Rizvi and Lingard 2009; Steiner-Khamsi 2014). Yet there has also been increasing critique of such cross-border policy learning due to incompatibility with local contexts (e.g. Coffield 2012; Feniger and Lefstein 2014). This tension is reflected in both the opening and closing pages of *World Class: How to Build a 21st-Century School System*, where Andreas Schleicher (2018) notes that international benchmarking of educational practices ‘is not about copying . . . solutions from other countries; it is about looking seriously and dispassionately at good practice in our own countries and elsewhere to become knowledgeable of what works in which contexts’ (pp. 14, 279). This statement sits uneasily with the fact that Schleicher is the architect of the OECD’s Programme for International Student Assessment (PISA) that itself fuels the popularity of borrowing ‘best practices’ from high-scoring countries (Fischman et al. 2019).

In this paper, I contribute to the case against two claims that disproportionately privilege the general over the specific: first, the assumption that teacher accountability is purely technical matter rather than a social and contextually embedded one; and, second, the use of international large-scale student assessments (ILSAs) such as PISA to generate universalised policy prescriptions. To highlight the role of context specificity in teacher accountability and in the patterns captured in ILSA data, I explore the extent to which the relationship between teacher accountability instruments and student outcomes depends on sociocultural context. This exploration uses multilevel modelling of cross-country surveys of education – including the OECD’s PISA data, which have been used to justify ‘best practice’ arguments (e.g. Mourshed, Krawitz, and Dorn 2017) – alongside cross-country surveys of cultural values. I show that the relationship between teacher accountability instruments and student outcomes is significantly moderated by sociocultural context, such that more extensive teacher accountability practices are associated with better student outcomes under certain sociocultural conditions, but with worse student outcomes in other settings. On one level, and to the extent that cross-sectional analyses of ILSAs are taken as sources of policy evidence, this finding challenges the assumption that teacher accountability is a technical rather than a social matter. On another level, this finding supports the case against the acontextual use of ILSAs for policy prescriptions because it demonstrates that the relationship between practices and outcomes in at least one policy area is contingent on a contextual feature that is not typically included in ILSA analyses.

In the next section, I summarise the literature on areas related to this study: the popularity and contentiousness of formal, performance-based teacher accountability reforms; the influence of context, especially sociocultural context, on teacher accountability; the generalisability of ILSA data; and the purposes for which statistical analyses have combined ILSA data with sociocultural survey data. Next, I present the data, methods, and results of the statistical analysis described above. I conclude by discussing the implications of this analysis for the interpretation of ILSA data and the design of teacher accountability policy.

## The context of this study

### *Varieties of teacher accountability*

Recent years have seen a growing emphasis on formally codified – and often efficiency-oriented, performance-based, consequence-linked, or ‘managerial’ – accountability structures in education policy (Tulowitzki 2016; Verger and Parcerisa 2017) and in public management more generally (Muller 2018; Pollitt and Bouckaert 2017). In 2015, routine teacher appraisals directly affected teachers’ pay in countries ranging from Chile to Hungary to Singapore (OECD, 2016b).

Nonetheless, there is considerable disagreement about whether such managerial forms of teacher accountability lead to better student outcomes. This disagreement is due partly to wide variation in conceptions of accountability, whether for teachers or for other actors (Broadfoot and Osborn 1993; Hopmann 2008; Koppell 2005; UNESCO 2017). Although the statistical constructs analysed in this paper reflect a narrower conception of teacher accountability instruments as formal and managerial (because they were limited to data available in PISA surveys), I define teacher accountability instruments broadly, drawing on two widely cited conceptions of accountability: Bovens’ (2007) definition of accountability as ‘a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgment, and the actor may face consequences’ (p. 450); and Romzek and Dubnick’s (1987) argument that ‘public administration accountability involves the means by which public agencies and their workers manage the diverse expectations generated within and outside the organization’ (p. 228). Accordingly, I define teacher accountability instruments as *tools, practices, and structures that aim to orient teacher practice toward stakeholder expectations by (a) collecting information about teachers’ individual or collective practice and communicating this information to stakeholders, (b) setting standards by which stakeholders judge teacher practice, and/or (c) allocating consequences based on stakeholders’ judgements of teachers’ practice*. This definition includes instruments targeting not only teachers’ individual actions but also their collective practice, since department- or school-level evaluation and incentives can affect teachers’ professional experiences (e.g. Ingersoll, Merrill, and May 2016). It also includes the informal, highly localised, and sometimes tacit forms of accountability that can be pivotal in teacher practice (e.g. Abelman and Elmore 1999).

While some scholars recommend wider use of performance-based teacher accountability (e.g. Hanushek 2019), others call for caution. Some critics argue that the value-added models often used in test-based accountability are statistically flawed and fundamentally unfair (e.g. Amrein-Beardsley 2014). Others critique certain approaches to teacher accountability for the implicit assumption that the binding constraint in student outcomes is teacher effort, when other factors may impose more significant constraints (e.g. Benveniste 1985; Wagner 1989). Still others emphasise the detrimental side effects of managerial, test-based accountability (Thiel, Schweizer, and Bellmann 2017; Zhao 2018), such as diverting resources toward test preparation (Altrichter and Kemethofer 2015; Booher-Jennings 2005) and compromising teachers’ mental health and professional identities (Holloway and Brass 2018; Liew 2012; Von der Embse et al. 2016).

### ***Teacher accountability, sociocultural context, and generalisability***

Another line of critique against extensive performance-based teacher accountability argues that the effect of any teacher accountability instrument is contingent on the context in which it is embedded. These arguments draw equally on case studies of educational accountability in practice (e.g. Broekman 2016; Narwana 2015; Mizel 2009), and on theory-of-change frameworks that emphasise the role of context in implementation (Cambridge Assessment 2017; McDonnell and Elmore 1987; and Monaghan and King 2018 in education policy; Andrews, Pritchett, and Woolcock 2017; Bates and Glennerster 2017; Cartwright and Hardie 2012; and Williams 2020 in public policy).

However, sociocultural context typically receives less systematic attention than other aspects of national-level context. For example, UNESCO's (2017) Global Education Monitoring (GEM) report on accountability emphasised economic, political, and social contexts in designing education accountability systems. However, the discussion of salient contextual characteristics covered resource levels, organisational structures, political institutions, and interest groups – but gave little attention to cultural and social patterns that may influence educational actors' responses to accountability structures. One reason for the GEM Report's inattention to national culture in educational accountability is that there has been relatively little research in this area (for an overview, see Hwa 2019, Chapter 2). Another reason is that analysing culture across countries is fraught with the risk of unfair generalisations that reinforce prejudices or facilitate morally objectionable conclusions (see, for example, Kim 1994; Sen 1999, on the 'Asian values' argument for non-democratic rule in Southeast Asia). However, established bodies of research in other fields (e.g. Alesina and Giuliano 2015, in economics; Hofstede 2001, in organisational studies; Markus and Kitayama 2010, in cross-cultural psychology; Pawson and Tilley 1997, in policy analysis) show that people's responses to the structures around them can be strongly influenced by sociocultural context. Furthermore, Gelfand, Lim, and Raver (2004) and Velayutham and Perera (2004) have argued that the workings of accountability vary by certain cultural traits.

Culture, like accountability, is a diffuse and contested concept. To gain some analytic traction, I focus on sociocultural context. My definition of sociocultural context draws on two sources. Firstly, Maxwell's (2012) realist-informed proposition that 'a culture is a *system* of individuals' conceptual/meaningful structures (minds) found in a given social system, and is not intrinsically shared, but participated in' (p. 28, emphasis original). Secondly, Markus and Kitayama's (2010) work in cultural psychology, in which they locate culture not in stable beliefs inside people, but in 'patterns of ideas, practices, institutions, products, and artifacts' (p. 422) situated in the world. Hence, I define sociocultural context as *dominant patterns of ideas and practices in a given social system that influence people's interactions with their environments*.

Thus, in exploring the extent to which the relationship between teacher accountability instruments and student outcomes is influenced by sociocultural context, I am, in effect, exploring whether the tools, practices, and structures that aim to orient teacher practice toward stakeholder expectations depend on dominant patterns of ideas and practices in

the given social system.<sup>1</sup> This posited mechanism has two elements: firstly, that accountability instruments orient teacher practice toward stakeholder expectations by shaping teachers' subjective perspectives and priorities; and, secondly, that teachers' subjective perspectives and priorities are themselves shaped by dominant sociocultural patterns. To give support for the first element, i.e. that teachers' subjective perspectives mediate between accountability instruments and teacher practice, Müller and Hernández's (2010) study of seven European countries found that teachers were largely sceptical about accountability instruments because these policies generated peripheral paperwork rather than enhancing the classroom teaching that they considered their chief responsibility (see also Abelmann and Elmore 1999; McLaughlin 1987; Spillane 2009 for theoretical arguments). To give support for the second element, i.e. that teachers' subjective perspectives are shaped by dominant sociocultural patterns, Broadfoot and Osborn (1993) found systematic differences between English and French teachers' perceptions of their professional responsibilities, which were influenced by cultural and ideological assumptions.

This interplay between accountability instruments, subjective perspectives, and socio-cultural context has been documented in a few settings. For example, a major teacher quality reform in Indonesia included plans to train teachers as peer evaluators, but teachers in this highly hierarchical sociocultural context questioned the authority of colleagues – rather than supervisors or headteachers – to evaluate their work (Broekman 2016). Similarly, teachers in India challenged certain community accountability structures because they regarded themselves as high-status professionals beyond the purview of low-status villagers (Narwana 2015).

In this paper, I add to arguments about the importance of sociocultural context in teacher accountability by drawing on empirical data from another hotly contested area of education policy: ILSAs and their associated surveys of classroom, household, and school characteristics. On one hand, ILSA results are often deployed by policymakers to legitimise their policy choices (Fischman et al. 2019). Claiming that a policy initiative mimics that of high-performing systems can be a persuasive political strategy (Andrews, Pritchett, and Woolcock 2017) – even if the legitimised policies diverge from the high-performing models that they purportedly mimic (You 2017). Besides providing legitimation, ILSAs can be attractive evidence sources because their samples span diverse institutional setups and education policy approaches, which are often quite homogeneous within countries (Hanushek and Woessmann 2011; Wagemaker 2010).

On the other hand, treating ILSAs as a compass for education policymaking can be problematic. Feniger and Lefstein (2014) show that the PISA 2009 scores of children who migrated from China to New Zealand and Australia were more similar to the scores of their former peers in China than their current schoolmates in Australasia (see also De Philippis and Rossi 2021; Jerrim 2015). They argue that this finding challenges the assumption in PISA-based policy reasoning that cross-country variability in PISA scores is caused by education system features at the time of the test. Although official PISA reports include clear statements that their findings are correlational (e.g. OECD 2020, 41–43), they are often interpreted as causal, thus fuelling the preoccupation with adopting 'best practices' from countries with high ILSA scores (for some critiques of this



phenomenon, see Coffield 2012; Sellar and Lingard 2013). While Feniger and Lefstein (2014) use PISA data to critique the ILSA-inspired ‘best practices’ approach based on faulty causality, in this paper I use ILSA data to critique this approach from another angle: that of limited generalisability.

### ***Combining cross-country survey data on education and sociocultural context***

To show the limited generalisability of the policy-performance association in ILSAs, I combine educational data from two ILSAs – PISA and TIMSS – with sociocultural data from two cross-country surveys of culture – the World Values Survey/European Values Study (WVS/EVS) and Hofstede’s Values Survey Module. While ILSA analyses often incorporate the national economic context, as proxied by GDP-related measures (see, e.g. OECD 2020, 43), fewer analyses include data on sociocultural context.

Still, country-level sociocultural indicators have been productively merged with ILSA data for at least three purposes. Firstly, West and Woessmann (2010) and Heller-Sahlgren (2018) use sociocultural data in casual identification; specifically, in instrumental-variable strategies that use historic data on the prevalence of Catholicism to provide exogenous variation in private school enrolment. Secondly, some ILSA analyses aim to identify correlations between country-level sociocultural characteristics and student achievement, with little agreement between studies about which sociocultural factors matter most (Benoliel and Berkovich 2018; He, van de Vijver, and Kulikova 2017; Meyer and Schiller 2013). Others have used national sociocultural data to examine the relationship between gender gaps in mathematics proficiency and societal gender norms (Fryer and Levitt 2010; Guiso et al. 2008; Rodríguez-Planas and Nollenberger 2018), or between students’ career ambitions and societal values (Han, Borgonovi, and Guerriero 2018).

Finally, cross-country datasets on education and culture can be combined to examine whether contextual variables interact with other predictors to moderate – that is, either intensify or attenuate – the relationship between educational inputs and outcomes. Using data from PISA and Hofstede’s survey, Chiu and Klassen (2010) find that some aspects of sociocultural context slightly moderate the relationship between students’ perceptions of themselves and their mathematics proficiency. Coco and Lagravinese (2014) find that the relationship between educational expenditure and PISA scores is moderated by cronyism, as measured in the WVS – suggesting that cronyism creates disincentives to acquire skills, thus reducing the efficiency of educational spending.

In this vein, I use WVS/EVS and Hofstede data to investigate how sociocultural context moderates the relationship between teacher accountability instruments and student achievement. While other studies have investigated whether the relationship between accountability instruments and student outcomes are moderated by other institutional features (e.g. school autonomy, as in Hanushek, Link, and Woessmann 2013; Woessmann 2016) or by country-level educational achievement (e.g. Bergbauer, Hanushek, and Woessmann 2018), this study is novel, to my knowledge, in investigating moderation effects from sociocultural context (see Hwa 2019, Chapter 2, for a literature search on this topic).



## Methods

### *Data and sampling*

To explore the relationship between teacher accountability instruments, sociocultural context, and student outcomes, I use publicly available secondary datasets on education and culture. In each regression, data from one educational survey is matched with GDP data from the Penn World Table and sociocultural data from the WVS/EVS, Hofstede's survey, or both. Missing values are excluded listwise.

The main ILSA dataset that I analyse is the 2015 wave of PISA, the OECD's Programme for International Student Assessment (OECD 2016a). The PISA 2015 dataset covers not only student-level proficiency scores for a nationally representative sample of school-going 15-year-olds, but also a wide range of contextual variables, including school-level questionnaire items on teacher accountability. In the main PISA 2015 analysis, I use a dataset which has no missing observations for any of the PISA variables of interest, nor for the WVS/EVS and Hofstede sociocultural scales. This dataset comprises 346,726 pupils from 12,764 schools across 57 countries. Additionally, I analyse data from PISA 2012 (OECD 2014b). In addition to offering an alternative dataset for sensitivity checks, the PISA 2012 questionnaires have a richer set of items on accountability than PISA 2015, as described below.

Besides PISA, I analyse data on eighth-grade pupils from the 2015 wave of the IEA's Trends in International Mathematics and Science Study (TIMSS; Martin et al. 2016). TIMSS 2015 questionnaires do not include enough accountability-related items to construct a measure of teacher accountability instruments, so I run analyses using student outcome and background data from TIMSS 2015 matched with country-level weighted means of the teacher accountability scales from PISA 2012 and 2015, in turn. The weights used to construct these country-level means were calculated PISA administrators to generate nationally representative statistics despite differences in school sizes, different numbers of schools sampled within each country, and different nonresponse rates (OECD 2017a). While the schools that participated in PISA 2015 may not be the same as those in TIMSS 2015 (nor those in PISA 2012), all of these datasets are nationally representative. Thus, when the accountability and student outcome data come from different assessment cycles, the accountability variables will only enter the model at the national level – similar to analyses combining student-level PISA data with national-level data on per capita GDP or the GINI coefficient (e.g. Condron 2011; Woessmann et al. 2009).

For national sociocultural context, I draw on two survey programmes. First, I use two waves of the World Values Survey (WVS), i.e. Wave 5 (conducted between 2005 and 2009; Inglehart et al. 2014a) and Wave 6 (2010–2014; Inglehart et al. 2014b), alongside one wave of the European Values Study (EVS), i.e. Wave 4 (2008–2010; EVS 2011). WVS/EVS is the largest international survey programme on culture, conducted as face-to-face interviews with nationally representative samples of at least 1,000 adult residents in each participating country per wave (EVS 2016; WVS Association n.d.). Additionally, I use two sociocultural indices – power distance and uncertainty avoidance – from Geert Hofstede's Values Survey Module. Hofstede's dataset is also known as the IBM study because the bulk of the surveying was conducted with IBM employees in 72 countries between 1967 and 1973 (Hofstede 2001). This longstanding research programme is

highly influential in cross-cultural survey measurement and organisational behaviour (see Taras, Kirkman, and Steel 2010, for a review). Besides sociocultural context, I also include country-level GDP data from the Penn World Table 9.0 (Feenstra, Inklaar, and Timmer 2016).

In matching the country-level contextual data to the ILSA datasets, I ensure appropriate time-ordering of predictor and outcome variables. For example, when analysing PISA 2012 student outcome data, I use sociocultural data that was collected prior to the PISA testing dates. (Note that PISA 2015 data were collected slightly later than TIMSS 2015 data, i.e. March to December 2015 for PISA, compared to October 2014 to May 2015 for TIMSS. Thus, for some countries, there may be a slight violation of time-ordering when I combine accountability data from PISA 2015 with outcome data from TIMSS 2015.) Matching country-level sociocultural data to the multilevel educational datasets requires the strong assumption that the survey-based indicators are adequate proxies for sociocultural context despite differing time lags between the sociocultural and educational surveys. Although I am reluctant to make any claims in principle about the stability of sociocultural context over time, this assumption about the adequacy of older sociocultural survey data as a proxy for more recent sociocultural context is not unusual in cross-country statistical analysis (e.g. Norris and Inglehart 2004; see also Hofstede 2001; Inglehart and Welzel 2005 for claims in support of this assumption).

The number of countries in each regression depends on the overlap in country participation between the educational, cultural, and economic datasets in question. Nonetheless, every regression includes both high- and low-performing education systems, spanning all six continents. For the full list of included countries, see Table S1 in the supplemental online material.

## **Operationalisation**

### ***Teacher accountability instruments***

To operationalise teacher accountability instruments, I draw on principal-reported data on teacher accountability instruments in PISA 2012 and 2015. I use two-parameter logistic item-response theory (IRT) modelling to construct scales for the extensiveness of teacher accountability instruments in any given school. These scales offer an overall snapshot of teacher accountability instruments, facilitating cross-country comparison. Additionally, using a single aggregate measure rather than multiple, correlated variables for each teacher accountability instrument facilitates model convergence when looking at country-level teacher accountability, given the limited country-level sample size.

IRT is likewise used by the OECD and the IEA to construct both student scores and contextual scales in PISA and TIMSS. A further benefit of IRT modelling is its capacity for generating a scale score for any case that has data on at least one of the items underlying the scale, thus yielding far less missingness in the generated scale variable. For example, 5.3% of schools in the pooled PISA 2012 and 2015 dataset did not have data on any of the accountability items, and hence do not have an IRT accountability score. In contrast, three times as many schools, 16.0%, did not have data on one or more items. Consequently, a scale construction method that could only generate scores for cases with

observations for all items, e.g. a simple count of the number of accountability instruments in a given school, would have excluded all 16.0% of those schools from the analysis.<sup>2</sup>

The scale draws on 21 principal-reported questionnaire items. Although they were chosen based on the broad definition of teacher accountability instruments above, most of the relevant items available carry formal, managerial connotations. These items fell into four categories: *how teachers are monitored* (4 items: student assessments, teacher peer review, lesson observations by school leaders, lesson observations by external persons); *what quality assurance approaches are used* (7 items: internal evaluation, external evaluation, written specification of educational goals, written specification of student performance standards, recording of attendance and professional development, recording of student outcomes, written feedback from students); *how student achievement data are shared* (3 items: posted publicly, tracked over time by an authority, provided directly to parents), and *what consequences might result from teacher appraisals* (7 items, PISA 2012 only: a change in salary, a financial bonus, professional development opportunities, a change in career advancement prospects, public recognition, a change in work responsibilities, a role in school development initiatives). For parameter estimates from the IRT model, see Table S2 in the supplemental online material. Notably, the parameter estimates suggest some construct validity in that the items with the highest difficulty parameters in Table S2 correspond to the most intensive or managerial forms of teacher accountability: public posting of student achievement data, teacher appraisal by external individuals, and financial rewards from teacher appraisals.

### **Student outcomes**

PISA and TIMSS assess multiple subjects, but individual students' proficiency scores are highly correlated across subjects. For simplicity, I focus on one subject per assessment wave. PISA includes questions on reading, mathematics, and science, but every wave assesses one of the three subjects in particular detail. I focus on the emphasised subject from each wave; i.e. science for PISA 2015 and mathematics for PISA 2012. TIMSS 2015 allocates equal coverage to mathematics and science, and I focus on mathematics. (To check the soundness of this decision to focus on one subject per assessment wave, I re-estimated the main regression model using the other PISA and TIMSS subjects as outcome variables. There were no differences in the direction or significance of key variables.)

### **National sociocultural context**

Given the multidimensionality of culture and society, I use two data sources on country-level sociocultural context, as noted above. From these datasets, I use proxy scales for social capital, power distance, and uncertainty avoidance. Although there are numerous other ways of conceptualising and measuring sociocultural differences (e.g. Green, Janmaat, and Han 2009; Inglehart and Welzel 2005; Markus and Conner 2013; Thompson, Ellis, and Wildavsky 1990), I focus on these constructs because they were identified, in a systematic literature search on teacher accountability and sociocultural context, as cultural patterns that are theoretically expected to moderate the effects of accountability instruments (Bryk and Schneider 2002; Cerna 2014; Iyengar 2012; Webber

2010, on social capital and trust; Broekman 2016; Gelfand, Lim, and Raver 2004; Narwana 2015; Velayutham and Perera 2004 on power distance and uncertainty avoidance; for the systematic literature search, see; Hwa 2019, Chapter 2).

First, from the WVS/EVS datasets, I use factor analysis to construct four scale variables for aspects of social capital, using country-level aggregate data published by survey administrators. I identify four sets of WVS/EVS questionnaire items that relate to social capital and are available in all three survey waves: *confidence in institutions* (12 items, e.g. parliament or labour unions; where each country's score on each item is the proportion of respondents stating that they have 'a great deal' or 'quite a lot' of confidence in the institution in question); membership in *civic networks* (7 items, e.g. religious organisations or recreational organisations, where each country's score on each item is the proportion of respondents stating that they belong to the voluntary activity in question); adherence to *civic norms* (4 items, e.g. cheating on taxes or avoiding a fare on public transport; where each country's score on each item is the weighted mean of a 10-point scale ranging from whether breaching the norm in question is 'always justifiable' to 'never justifiable'); and *social trust* (2 items; where each country's score on the first item is the proportion answering 'most people can be trusted' rather than 'you need to be very careful'; and on the second item is the weighted mean of a 10-point scale ranging from 'most people would try to take advantage of you if they got a chance' to 'most people would try to be fair'). All four WVS/EVS scales are coded such that higher values correspond to more social capital, i.e. greater confidence in institutions, more extensive membership in civic networks, stronger adherence to civic norms, and more social trust. For detailed questionnaire items and factor loadings of these sociocultural scales, see Tables S3 to S6 in the supplemental online material.

Alongside these newly constructed WVS/EVS factor variables, I include two preexisting sociocultural scales from Hofstede's IBM dataset. On the first scale, *power distance*, higher values correspond to a greater acceptance of hierarchical distributions of power. Higher values on the second scale, *uncertainty avoidance*, indicate a greater tendency toward anxiety and stronger preferences for stability. Hofstede calculates the power distance and uncertainty avoidance indices through linear combinations of country-level average responses to the pertinent questionnaire items (Hofstede 2001, 86, 150). For this analysis, I standardise power distance and uncertainty avoidance scores to correspond to the mean and spread of the social capital factor scores.

### **Control variables**

I use a relatively parsimonious set of controls, with one control variable at each level of analysis. My aim is not to capture as much variability in student achievement as possible; but rather to test whether sociocultural context does, in fact, moderate the relationship between teacher accountability instruments and student achievement. Accordingly, control variables are included only if there are theoretical or empirical grounds for expecting them to affect the relationship between teacher accountability, sociocultural context, and student outcomes.

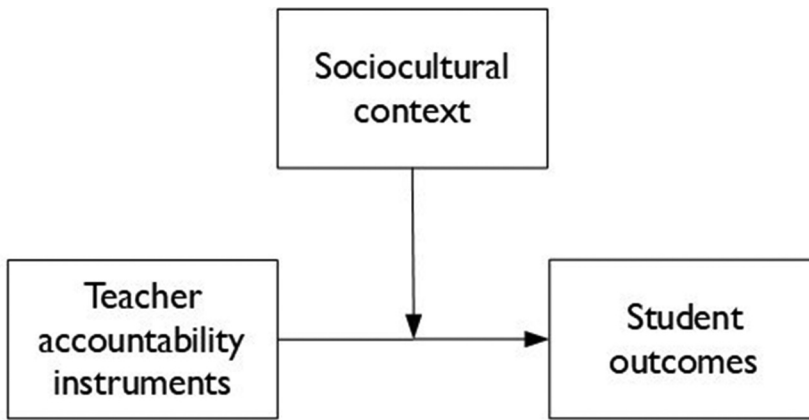
At the student level, I control for socioeconomic background, using the PISA scale for economic, social, and cultural status (ESCS) or the TIMSS scale for home educational resources, respectively; because less-privileged students are often concentrated in lower-performing schools that may be affected differently by teacher accountability instruments

(Diamond and Spillane 2004). At the teacher level, in the TIMSS regressions, I control for years of teaching experience, because teaching experience has been associated with different levels of student outcomes (Murnane and Phillips 1981; Hanushek and Rivkin 2006). (Teaching experience is not included in the PISA regressions because teacher-level questionnaires are not compulsory in PISA.) At the school level, I control for school capacity for responding to accountability incentives, because teacher accountability instruments may be more effective when schools have more freedom to change their practices. For PISA, I use the degree of school autonomy in decision-making (see Woessmann 2016 for an empirical analysis related to accountability and school autonomy). TIMSS 2015 did not include questionnaire items on decision-making autonomy, so instead I use a principal-reported TIMSS scale for the degree to which instructional capacity is constrained by inadequate resources (see Mbiti et al. 2019 for a randomised-control trial related to accountability and resources). Finally, at the country level, I control for per-capita GDP for the year preceding the ILSA in question. The GDP variable was scaled in 2011 US\$10,000s and centred at \$30,000, to give it an order of magnitude similar to that of the country-level sociocultural variables while remaining meaningfully interpretable. Although I do not subscribe to modernisation theories that associate developed countries with ‘modern’ values and developing countries with ‘traditional’ values (e.g. Inglehart and Welzel 2005), it is empirically true in my dataset that GDP is moderately correlated with some of the sociocultural constructs. Hence, to forestall the spurious attribution of moderation from national resource levels to moderation from national culture, I interact teacher accountability not only with the sociocultural scales but also with GDP.

Summary statistics for each variable are available in Table S7 of the supplemental online material.

### **Modelling**

To address the question of *to what extent does the influence of teacher accountability instruments on student outcomes depend on sociocultural context*, this statistical analysis is framed as a multilevel moderation model. That is, I posit that the relationship between teacher accountability instruments and student outcomes can either be intensified or weakened by sociocultural context; thus, sociocultural context moderates the relationship between accountability instruments and student outcomes, as illustrated in (Figure 1). Although both (Figure 1) and the language in this paper present the association between these three constructs as (sociocultural) context moderating the relationship between teacher accountability and student outcomes, it is important to note that the analysis uses cross-sectional datasets that preclude causal claims – such that it would also be plausible to interpret the analysis as accountability instruments moderating the relationship between context and outcomes. In practice, context and policy are mutually influencing, and they jointly influence outcomes. However, I frame context as the moderator because one goal of this paper is to contribute to the critique of acontextual policymaking in education. If the aspiration is improving student outcomes, then changing teacher accountability instruments to better suit the sociocultural context is a more viable path, at least in the short term, than attempting to reshape the wider sociocultural context to the mould of preexisting teacher accountability instruments.



**Figure 1.** Posited model, in which sociocultural context moderates the relationship between teacher accountability instruments and student outcomes.

I use statistical techniques that are appropriate for the sampling and assessment designs of the educational surveys (Jerrim et al. 2017; Martin, Mullis, and Hooper 2016; OECD 2014a, 2017a; Rutkowski et al. 2010). Multilevel modelling accounts for the clustering of students into classes and/or schools in PISA and TIMSS. School-level weights and conditional student- and teacher-level weights account for varied response rates as well as oversampling of certain strata (e.g. certain regions or certain school types) within countries. Additionally, the PISA and TIMSS datasets include several plausible values (10 plausible values in PISA 2015, and 5 in TIMSS 2015 and PISA 2012) for each student proficiency score. These plausible values are random draws from IRT-based probability distributions of each student's 'true' score. I rerun every regression for each plausible value, and then combine the coefficient and standard error estimates using Rubin's rules (Laukaityte and Wiberg 2017; OECD 2009; Rubin 1996). All regressions are estimated in MLwiN 3.0.2 using its iterated generalised least squares procedure, a form of full-information maximum likelihood estimation. The number of levels in each model depends on the sampling design of the respective educational dataset: PISA models have three levels (pupil, school, country); TIMSS models have four (pupil, teacher, school, country). All regressions use sandwich estimators for standard errors to mitigate the effects of potential heteroskedasticity.

Since I am interested in the effects of teacher accountability instruments both between countries and within countries, the PISA models include terms for both the country-level weighted mean of the school-level IRT estimate for teacher accountability as well as the difference between each school's score and the respective country mean. This is sometimes called a within-between model (e.g. Bell and Jones 2015), since the country-level mean measures variation between countries, whereas the school-level differential measures variation within countries. However, as noted above, TIMSS models do not have matched school-level data on accountability, so they only have a country-level weighted mean taken from PISA.

Where possible, I estimate two sets of regressions for each dataset: regressions that (a) include all six sociocultural constructs and their associated interaction terms; and (b) include just one sociocultural construct and its associated interaction term, thus covering all six constructs in a series of six regressions. While (a) is analytically preferable because it accounts for the interplay between the sociocultural constructs, it is not always empirically feasible. Due to the smaller number of countries in the TIMSS datasets, TIMSS regressions following option (a) showed indications of multicollinearity. Since such indications did not appear in TIMSS regressions following option (b), it is likely that the multicollinearity in (a) is due to over-fitting, i.e. including too many country-level variables in a regression with too few country cases. Accordingly, I present results from (a) where possible, but show results from (b) when (a) appears over-fitted.

For the main model with PISA 2015 data, I estimate:

$$\begin{aligned}
 Proficiency_{psc} = & \beta_0 + \beta_1 ESCS_{psc} + \beta_2 Autonomy_{sc} + \beta_3 GDP_c + \\
 & \beta_4 AccountabilityDiff_{sc} + \beta_5 Accountability_c + \\
 & \beta_6 AccountabilityDiff_{sc} * ESCS_{psc} + \beta_7 AccountabilityDiff_{sc} * Autonomy_{sc} + \\
 & \beta_8 AccountabilityDiff_{sc} * GDP_c + \beta_9 Accountability_c * ESCS_{psc} + \\
 & \beta_{10} Accountability_c * Autonomy_{sc} + \beta_{11} Accountability_c * GDP_c + \\
 & \beta_{12} Sociocultural_c + \beta_{13} AccountabilityDiff_{sc} * Sociocultural_c + \\
 & \beta_{14} Accountability_c * Sociocultural_c + v_c + u_{sc} + e_{psc}
 \end{aligned}$$

where  $Proficiency_{psc}$  is the proficiency score of pupil  $p$  in school  $s$  in country  $c$ . Control variables comprise pupil economic, social, and cultural status ( $ESCS_{psc}$ ), school autonomy ( $Autonomy_{sc}$ ), and national per-capita GDP ( $GDP_c$ ). The main explanatory variables are  $Accountability_c$ , the country-level weighted mean of teacher accountability, and  $AccountabilityDiff_{sc}$ , the school-level teacher accountability differential (i.e. the difference between each school's teacher accountability score and the  $Accountability_c$  for the relevant country).  $Sociocultural_c$  represents a vector of the six sociocultural constructs. I also include interactions between each of the two accountability variables and each of the other explanatory variables, including the controls. The latter are included to ensure that I do not erroneously attribute moderation effects to sociocultural context when those effects instead result from other contextual characteristics that may be correlated with the sociocultural constructs, such as GDP. Finally,  $v_c$ ,  $u_{sc}$ , and  $e_{psc}$  are error terms at each level. Models using other datasets or subsamples have the same overall form, with some deviations for TIMSS models as noted above. All models were estimated with random intercepts but fixed slopes between groups.

To assess robustness, I run a range of sensitivity checks. First, I test the model using data from PISA 2015, PISA 2012, and TIMSS 2015. I use the PISA 2012 data in two ways: simply repeating the PISA 2015 analysis with PISA 2012 data, to check the robustness of the models across assessment waves; and matching student outcome data from PISA 2015 with country-level accountability data from PISA 2012, to account for possible time lags in the effect of accountability instruments. I also estimate the model for two separate TIMSS datasets, matched with country-level accountability data from PISA 2015 and PISA 2012, respectively. Additionally, I analyse subsamples of the PISA 2015 and 2012 data containing observations only from OECD countries, as well as a subset of PISA 2015



data containing observations only from publicly funded schools. Finally, for models that had significant interactions between accountability and sociocultural context, I re-estimated the regressions with dummy variables for outlying countries (e.g. Vietnam and China in the main PISA 2015 regression, as identified in residual plots). The inclusion of these country dummies did not materially affect either the magnitude or the significance of the interaction terms of interest.

## Results

(Table 1) summarises the significance and direction of parameter estimates for the interaction between country-level teacher accountability and each of the six sociocultural constructs, across all the models. These models account for a substantial proportion – ranging from 63% to 79% – of the between-country variance in each dataset. For example, for the PISA 2015 full sample, moving from a null model with no explanatory variables to the full model summarised in (Table 1) reduced the unexplained country-level variance in the dataset from 1980.19 to 500.19 (a 75% reduction), as shown in Table S8 of the supplemental online material.

As shown in (Table 1), the interaction between  $Accountability_c$  and civic norms is consistently negative and significant for all the PISA 2015, PISA 2012, and TIMSS 2015 models tested. None of the other sociocultural constructs consistently moderated the relationship between teacher accountability instruments and student outcomes.<sup>3</sup> Additionally, none of the interactions (sociocultural or otherwise) with the school-level accountability differential,  $AccountabilityDiff_{sc}$ , are significant. These school-level interactions and the rest of the results for the PISA 2015 full sample, presented as a series of nested models, are available in Table S8.

For a more detailed look, (Table 2) shows parameter estimates from the sensitivity checks for the main effect of  $Accountability_c$  and its interactions with civic norms and the non-sociocultural moderators. The unmoderated effect of  $Accountability_c$  on student proficiency is insignificant in all of the models. However, many of the interactions

**Table 1.** Summary of sociocultural constructs that significantly moderate the relationship between  $Accountability_c$  and pupil test scores.

	Models with all six sociocultural constructs entered simultaneously				Models with each sociocultural construct entered singly			
	PISA 2015 science				PISA 2012 maths		TIMSS 2015 maths	
	Full sample	OECD countries	Public schools	Accountability from 2012	Full sample	OECD countries	Accountability from 2015	Accountability from 2012
Confidence in institutions								
Civic networks				--		-		
Civic norms	--	--	--	-	-	-	--	--
Social trust				++		+	+	
Power distance								
Uncertainty avoidance						+		
N (countries)	57	36	56	54	52	36	23	22

1 symbol (+/-) indicates  $p < .05$ ; 2 symbols (+ +/- -) indicate  $p < .01$ . In the TIMSS models, each sociocultural construct was entered singly, in a series of six separate regressions, because the smaller country sample size led to overfitting when all six constructs and their interactions were included simultaneously. Full results can be provided upon request.

**Table 2.** Parameter estimates for the direct and moderated associations between pupil test scores and *Accountability<sub>c</sub>* (selected variables only).

<b>PISA 2015</b>		<i>With all six sociocultural constructs in the model</i>				
<i>Y<sub>psc</sub></i> = science proficiency	Full sample	OECD only		Public schools only	With <i>Accountability</i> from PISA 2012	
<i>Accountability<sub>c</sub></i>	0.03 (15.87)	-22.02 (18.45)		2.80 (17.13)	-5.50 (10.19)	
<i>Accountability<sub>c</sub></i> * <i>ESCS<sub>psc</sub></i>	-5.74* (2.50)	-2.29 (3.03)		-3.76 (2.29)	-7.16** (1.81)	
<i>Accountability<sub>c</sub></i> * <i>School autonomy<sub>sc</sub></i>	-4.56 (12.11)	20.79 (16.59)		2.75 (10.57)	-6.99 (7.43)	
<i>Accountability<sub>c</sub></i> * <i>GDP<sub>c</sub></i>	9.85* (4.59)	36.48** (12.99)		5.24 (5.19)	8.65* (3.66)	
<i>Accountability<sub>c</sub></i> * <i>Civic norms<sub>c</sub></i>	-28.74** (10.92)	-67.98** (14.13)		-35.56** (12.56)	-26.16* (11.29)	
N Pupils	346,726	210,533		272,204	340,680	
Schools	12,764	8,064		99,95	12,510	
Countries	57	36		56	54	
<b>PISA 2012</b>		<i>With all six sociocultural constructs in the model</i>				
<i>Y<sub>psc</sub></i> = mathematics proficiency	Full sample	OECD only				
<i>Accountability<sub>c</sub></i>	-24.98 (16.59)	-21.12 (19.43)				
<i>Accountability<sub>c</sub></i> * <i>ESCS<sub>psc</sub></i>	-6.57** (2.22)	-5.50 (4.41)				
<i>Accountability<sub>c</sub></i> * <i>School autonomy<sub>sc</sub></i>	-5.51* (2.29)	-5.66 (4.21)				
<i>Accountability<sub>c</sub></i> * <i>GDP<sub>c</sub></i>	12.69 (7.79)	21.56** (8.04)				
<i>Accountability<sub>c</sub></i> * <i>Civic norms<sub>c</sub></i>	-16.52* (8.10)	-29.39* (12.02)				
N Pupils	375,207	275,715				
Schools	14,840	11,169				
Countries	52	36				
<b>TIMSS 2015</b>		<i>With only civic norms in the model</i>				
<i>Y<sub>ptsc</sub></i> = mathematics proficiency	With <i>Accountability</i> from PISA 2015	With <i>Accountability</i> from PISA 2012				
<i>Accountability<sub>c</sub></i>	26.59 (15.98)	-3.17 (13.93)				
<i>Accountability<sub>c</sub></i> * <i>Home resources<sub>ptsc</sub></i>	-12.59** (3.67)	-11.99** (2.75)				
<i>Accountability<sub>c</sub></i> * <i>Teaching experience<sub>ts<sub>c</sub></sub></i>	0.13 (0.11)	0.06 (0.11)				
<i>Accountability<sub>c</sub></i> * <i>School resources<sub>sc</sub></i>	-2.20 (4.23)	-6.83 (3.91)				
<i>Accountability<sub>c</sub></i> * <i>GDP<sub>c</sub></i>	21.57* (10.55)	23.59** (8.58)				
<i>Accountability<sub>c</sub></i> * <i>Civic norms<sub>c</sub></i>	-71.20** (16.48)	-40.39* (16.96)				
N Pupils	118,363	120,117				
Schools	6,147	6,062				
Teachers	3,761	3,779				
Countries	23	22				

ESCS = Economic, Social, and Cultural status. Full results for the PISA 2015 full sample are available in Table S8 of the supplemental online material. Full results for other models can be provided upon request.

\**p* < 0.05. \*\**p* < 0.01.

between *Accountability<sub>c</sub>* and the contextual moderators are significant, most consistently with civic norms. The consistency of the significant interaction between *Accountability<sub>c</sub>* and civic norms across different test cycles and subsamples of the data suggests that this aspect of national sociocultural context does, in fact, moderate the relationship between student outcomes and teacher accountability instruments.

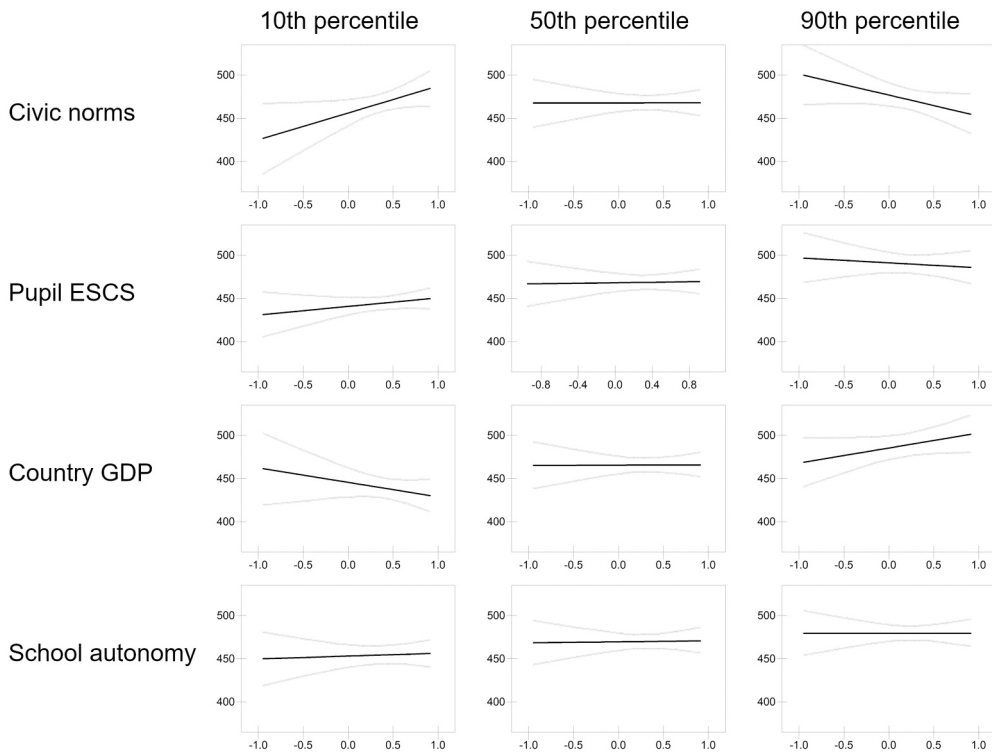
In summary, out of the six sociocultural constructs that were theoretically expected to affect the relationship between teacher accountability instruments and student outcomes, the regressions presented in this section only found evidence for moderation by one of these six constructs, i.e. civic norms. These civic norms interaction terms were robust

across different assessment cycles and subsamples of the data. Crucially, these moderation effects are present despite controlling for interactions between Accountability<sub>c</sub> and pupil socioeconomic status, school autonomy, and per-capita GDP.

To visualise the implications of the significant interaction terms, (Figure 2) illustrates these moderation relationships for the full-sample PISA 2015 model by showing predicted science proficiency scores against Accountability<sub>c</sub> at the 10th, 50th, and 90th percentiles of different contextual moderators, with all other variables held constant at their means. Thus, these predicted scores incorporate all of the parameter estimates the full regression model, including both the non-interacted and interacted terms; as well as the empirical values of each explanatory variable, represented by the range of Accountability<sub>c</sub>, the various percentiles of the contextual moderator of interest in each row, and the mean values of all the other variables in the dataset. The first three rows of the figure show predictions for different levels of the three contextual variables that significantly moderated the relationship between Accountability<sub>c</sub> and science proficiency, i.e. civic norms, ESCS, and GDP (ordered from the largest p-value on the interaction term to the smallest). I also include school autonomy, as an example of a contextual variable that did not significantly interact with Accountability<sub>c</sub>.<sup>4</sup>

From the variety of slopes in these graphs, it is evident that teacher accountability instruments can have a positive (upward-sloping), negative (downward-sloping), or negligible (flat) overall effect on student outcomes, depending on context. For example, in the civic norms row, the leftmost graph reflects the model prediction that pupil science proficiency scores will increase as Accountability<sub>c</sub> increases, for a hypothetical country with a civic norms score at the 10th percentile. However, for a country at the 50th percentile of civic norms, increasing levels of Accountability<sub>c</sub> have no effect of predicted science proficiency. At the 90th percentile of civic norms, pupil science proficiency is expected to decrease as Accountability<sub>c</sub> increases. In contrast, the relationship between Accountability<sub>c</sub> and science proficiency is not affected by school autonomy given that the predicted score plots are similarly flat for all three levels of school autonomy. (However, school autonomy does have a direct, unmoderated association on pupil science proficiency, as indicated by the different intercepts in the three plots.)

The interaction between teacher accountability and sociocultural context can affect the association between these variables and student outcomes considerably. Comparing the rightmost ends of each plot – that is, the 100th percentile of Accountability<sub>c</sub> – a country at the 10th percentile of civic norms adherence would be expected to outscore a country at the 90th percentile of civic norms by 30 points, with all other variables held constant at their means. (Empirically, the 100th percentile of Accountability<sub>c</sub> corresponds to Russia, where the typical school had 13 out of the 14 the teacher accountability instruments in PISA 2015 questionnaires.) Conversely, at the 0th percentile of Accountability<sub>c</sub> at the leftmost ends of each plot (which corresponds to Greece, where the typical school had 8 of the 14 teacher accountability instruments), a country at the 90th percentile of civic norms would outscore a country at the 10th percentile of civic norms by 73 points.



**Figure 2.** Predicted PISA 2015 science proficiency scores (and 95% confidence intervals) against Accountability<sub>c</sub> for the 10th, 50th, and 90th percentiles of each contextual moderator. Note. All predictions are based on the regression with the full sample of PISA 2015 data. Each row shows predicted science proficiency scores against Accountability<sub>c</sub> for the 10th, 50th, and 90th percentile of the named contextual predictor. All other variables are held constant at their means.

Given that the PISA 2015 results report interprets 30 score points as being approximately equal to one year of schooling (OECD 2016a, 65), these differences are substantial.

## Discussion

To the extent that we want to use cross-sectional analyses of ILSAs as sources of policy evidence, the results of this analysis suggest that we cannot responsibly generalise from these assessments to the point of overlooking the specificities of national sociocultural context. Although the interaction between country-level teacher accountability and national sociocultural context was only significant for one of the six sociocultural constructs, these results were consistent across different subsamples of PISA 2015, PISA 2012, and TIMSS 2015 data. Moreover, the magnitude of these effects was large.

However, the upshot of this analysis is *not* that we should conduct more ILSA analyses with more sociocultural correlates. One reason to be cautious about such analyses is the sheer volume of assumptions about sample representativeness and modelling – as

described throughout this paper – that are required to merge and analyse educational and sociocultural datasets. Moreover, cross-country sociocultural metrics have at least as many limitations as their educational counterparts, not least in cross-cultural comparability (for an overview, see Survey Research Center 2016.) Even the most rigorous measures – such as higher-quality self-report questionnaire data in the Global Preference Survey (analysed alongside PISA data in Hanushek et al. 2020) or observational data as in the 40-country ‘lost wallet’ experiment (Cohn et al. 2019) – are vulnerable to the fact that any standardised benchmarking of sociocultural context vastly reduces the complexity, interactivity, and narratives that are embedded in any given setting. In other words, the tension between the general and the specific, discussed in the introduction with respect to teacher accountability policy and cross-country policy learning, certainly applies to the analysis of sociocultural context. Thus, any such statistical analysis runs the risk of false negatives because of variable selection.

For example, had I used only the Hofstede scales in this analysis, I may have erroneously concluded that sociocultural context was not related to the efficacy of teacher accountability instruments for student performance in PISA and TIMSS. Additionally, one reason why civic norms emerged as a significant moderator of the relationship between teacher accountability instruments and student outcomes may be due to the questionnaire items underlying the sociocultural scales. While some items for other sociocultural constructs asked about respondents’ actions (whether they were members of certain civic networks) and some asked about their beliefs (whether they were confident in institutions, trusted other people, tolerated power differentials, or preferred stability over uncertainty), the civic norms items addressed both areas, i.e. whether respondents believed certain actions could be justified. Thus, the civic norms scale relates to why people do what they do – thus capturing an aspect of motivation, which is both socioculturally contingent and pivotal to the relationship between accountability and outcomes (Wagner 1989; UNESCO 2017; Hwa 2021).<sup>5</sup> Yet the inclusion of a motivation-related scale in this analysis was a lucky coincidence, as the WVS/EVS scales were chosen solely as proxies for social capital, as described above.

Furthermore, this analysis does not demonstrate that trust is irrelevant to teacher accountability. It merely shows that the particular proxy that I used for social trust does not moderate the relationship between a set of ILSA proxies for teacher accountability and student learning. In a related interview study, I show that Finland’s and Singapore’s contrasting but comparably effective approaches to teacher accountability are closely related to how trust is distributed across different actors in each education system (Hwa 2021) – a multidimensional contextual feature that could hardly be encapsulated by the generic measure of social trust used in this analysis.

This challenge of reduction-in-standardisation applies to measures of context from cross-country surveys more generally. However detailed an ILSA background questionnaire may be, it only captures a small slice of the contextual features that affect education. One clear manifestation of this is the fact that the OECD and the IEA choose different sets of contextual items to include in each cycle; such that, for example, the PISA 2018 school questionnaire only included 10 out of the 21 items used to construct the teacher accountability instrument scale in this analysis (OECD, 2017b). The necessarily limited range of variables in survey datasets is one of several reasons – alongside the correlational rather than causal nature of cross-sectional ILSA analyses – why the main finding of this

analysis about adherence to civic norms moderating the effects of teacher accountability instruments should not be interpreted as a prescriptive guide for designing teacher accountability policy.

Rather, the results of this analysis instead support the case for viewing ILSA data and teacher accountability policy with circumspection, holding in balance both their limitations and their affordances. ILSAs can be highly valuable as broad-brush benchmarks of student learning, especially when they can expose grievous gaps in student learning (e.g. Pritchett and Viarengo 2021) or in settings where researchers cannot typically access representative student-level assessment data (as in my home country, Malaysia). However, except in the case of methodologically sophisticated quasi-experimental approaches, the typical cross-country, cross-sectional ILSA analysis is descriptive, not explanatory. Moreover, explanation of cross-country patterns in education policy and student outcomes should always require a description of the posited causal chain and the context – unlike a box in *PISA 2018 Results (Volume V)*, titled ‘What are the characteristics common to successful education systems?’, which notes that ‘there is no silver bullet in education’, but nonetheless offers a laundry list of characteristics identified correlationally in the dataset (OECD 2020, Box V.9.2, pp. 201–202).

As for teacher accountability, the analysis in this paper suggests that teacher accountability instruments interact with the contexts in which they are embedded – thus echoing a wide range of qualitative studies that reach similar conclusions from much more nuanced data (see Hwa 2019, Chapter 2 for some examples). Fundamentally, teacher accountability instruments should be a means, not an end. They should be a means of aligning teachers’ priorities with the shared goal of serving children throughout the education system. Such questions of priorities and motivation are inextricably embedded in the sociocultural context – especially for an endeavour as complex as education (Czerniawski 2011; Honig and Pritchett 2019; Hopmann 2008). Accordingly, recommendations for teacher accountability policy must suit the education system in question, rather than taking the form of blanket ‘best practices’. More broadly, both teacher accountability instruments and international student assessments are – or, at least, should be – valued to the extent that they help us to collectively improve the educational experiences and opportunities of all children.

## Notes

1. There is also a related question of the degree to which these reorientations of teacher practice improve student outcomes, but my working assumption for the purposes of this paper is that the most salient levels of context shaping the pathway from teacher practice to student outcomes are the classroom, community, and household, rather than national sociocultural context.
2. There did not appear to be systematic differences between the cases that had data on at least some accountability items – and, hence, were included in the analysis – and the 5.3% of cases that were excluded because they did not have data on any accountability items.
3. I also ran models with each sociocultural construct entered singly for each cut of the PISA 2015 and 2012 data, which are not shown in the table. Results were broadly consistent with those shown in (Table 1), with four exceptions, one of which related to civic norms: for the PISA 2012 full sample, the interaction between Accountability<sub>c</sub> and civic norms was insignificant, unlike in the model with all six sociocultural constructs. However, it was similarly negative in direction.

4. An additional set of graphs showing predicted scores across all combinations of the 10th, 50th, and 90th percentiles each for civic norms, ESCS, and GDP is available in Figure S1 in the supplemental online material.
5. A related concept is legal scholar Lynn Stout's (2010) 'unselfish prosocial behavior' (p. 99). Stout argues that unselfish prosocial behaviour can be eroded by an overemphasis on material reward and punishment because this sends the signal that it is appropriate to make decisions based on material and other selfish factors, thus crowding out altruistic justifications. This aligns with the association suggested in (Figure 2), i.e. that more managerial accountability instruments may backfire in contexts with strong civic norms.

## Acknowledgments

I am grateful to Panayiotis Antoniou, Ricardo Sabates, Geoff Hayward, William C. Smith, Jason Silberstein, and three anonymous reviewers, as well as participants at the EARLI SIG 18 & 23 conference in 2018 and CIES in 2019, for valuable feedback on different iterations of this analysis.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

The research in this paper was conducted as part of a PhD funded by the Gates Cambridge Trust (BMGF OPP1144).

## Notes on contributor

*Yue-Yi Hwa* is a research fellow at the Research on Improving Systems of Education (RISE) Programme at the Blavatnik School of Government, University of Oxford.

## ORCID

Yue-Yi Hwa  <http://orcid.org/0000-0002-4770-4940>

## References

- Abelmann, Charles, and Richard F. Elmore. 1999. "When Accountability Knocks, Will Anyone Answer?" CPRE Research Report Series RR-42. Philadelphia: Consortium for Policy Research in Education. Accessed 19 August 2017. [http://www.cpre.org/sites/default/files/researchreport/782\\_rr42.pdf](http://www.cpre.org/sites/default/files/researchreport/782_rr42.pdf)
- Alesina, Alberto, and Paola Giuliano. 2015. "Culture and Institutions." *Journal of Economic Literature* 53 (4): 898–944. doi:10.1257/jel.53.4.898.
- Altrichter, H., and D. Kemethofer. 2015. "Does Accountability Pressure through School Inspections Promote School Improvement?" *School Effectiveness and School Improvement* 26 (1): 32–56. doi:10.1080/09243453.2014.927369.
- Amrein-Beardsley, Audrey. 2014. *Rethinking Value-Added Models in Education: Critical Perspectives on Tests and Assessment-Based Accountability*. New York; London: Routledge.
- Andrews, Matt, Lant Pritchett, and Michael Woolcock. 2017. *Building State Capability: Evidence, Analysis, Action*. Oxford, New York: Oxford University Press.



- Bates, Mary Ann, and Rachel Glennerster. 2017. "The Generalizability Puzzle (SSIR)." *Stanford Social Innovation Review* Summer. Accessed 3 December 2018. [https://ssir.org/articles/entry/the\\_generalizability\\_puzzle](https://ssir.org/articles/entry/the_generalizability_puzzle)
- Bell, Andrew, and Kelvyn Jones. 2015. "Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data." *Political Science Research and Methods* 3 (1): 133–153. doi:10.1017/psrm.2014.7.
- Benotiel, Pascale, and Izhak Berkovich. 2018. "A Cross-National Examination of the Effect of the Schwartz Cultural Dimensions on PISA Performance Assessments." *Social Indicators Research* 139 (2): 825–845. doi:10.1007/s11205-017-1732-z.
- Benveniste, Guy. 1985. "The Design of School Accountability Systems." *Educational Evaluation and Policy Analysis* 7 (3): 261–279. doi:10.3102/01623737007003261.
- Bergbauer, Annika B, Eric A Hanushek, and Ludger Woessmann. 2018. "Testing." *NBER Working Paper* 24836. National Bureau of Economic Research. doi:10.3386/w24836.
- Booher-Jennings, Jennifer. 2005. "Below the Bubble: "Educational Triage" and the Texas Accountability System." *American Educational Research Journal* 42 (2): 231–268. doi:10.3102/00028312042002231.
- Bovens, Mark. 2007. "Analysing and Assessing Accountability: A Conceptual Framework." *European Law Journal* 13 (4): 447–468. doi:10.1111/j.1468-0386.2007.00378.x.
- Broadfoot, Patricia, and Marilyn Osborn. 1993. *Perceptions of Teaching: Primary School Teachers in England and France*. Cassell Education. London; New York: Cassell.
- Broekman, Art. 2016. "The Effects of Accountability: A Case Study from Indonesia." In *Flip the System: Changing Education from the Ground Up*, edited by Jelmer Evers and René Kneyber, 72–96. Oxford, New York: Routledge. doi:10.4324/9781315678573.
- Bryk, Anthony S., and Barbara Schneider. 2002. *Trust in Schools: A Core Resource for Improvement*. New York: Russell Sage Foundation.
- Cambridge Assessment. 2017. "A Cambridge Approach to Improving Education" Cambridge, United Kingdom: University of Cambridge Local Examinations Syndicate. Accessed 21 December 2018. <https://www.cambridgeassessment.org.uk/Images/cambridge-approach-to-improving-education.pdf>
- Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford; New York: Oxford University Press. Accessed 9 November 2018. doi:10.1093/acprof:osobl/9780199841608.001.0001
- Cerna, Lucie. 2014. "Trust: What It Is and Why It Matters for Governance and Education" *OECD Education Working Papers*. Accessed 16 November 2016. Paris: OECD. <http://www.oecd-ilibrary.org/content/workingpaper/5jxswcg0t6wl-en>
- Chiu, Ming Ming, and Robert M. Klassen. 2010. "Relations of Mathematics Self-Concept and Its Calibration with Mathematics Achievement: Cultural Differences among Fifteen-Year-Olds in 34 Countries." *Learning and Instruction* 20 (1): 2–17. doi:10.1016/j.learninstruc.2008.11.002.
- Coco, Giuseppe, and Raffaele Lagravinese. 2014. "Cronyism and Education Performance." *Economic Modelling* 38 (Supplement C): 443–450. doi:10.1016/j.econmod.2014.01.027.
- Coffield, Frank. 2012. "Why the McKinsey Reports Will Not Improve School Systems." *Journal of Education Policy* 27 (1): 131–149. doi:10.1080/02680939.2011.623243.
- Cohn, Alain, Michel André Maréchal, David Tannenbaum, and Christian Lukas Zünd. 2019. "Civic Honesty around the Globe." *Science* June: eaau8712. doi:10.1126/science.aau8712.
- Condron, Dennis J. 2011. "Egalitarianism and Educational Excellence Compatible Goals for Affluent Societies?." *Educational Researcher* 40 (2): 47–55. doi:10.3102/0013189X11401021.
- Czerniawski, G. 2011. "Emerging Teachers-Emerging Identities: Trust and Accountability in the Construction of Newly Qualified Teachers in Norway, Germany, and England." *European Journal of Teacher Education* 34 (4): 431–447. doi:10.1080/02619768.2011.587114.
- De Philippis, Marta, and Federico Rossi. 2021. "Parents, Schools and Human Capital Differences across Countries." *Journal of the European Economic Association* 19 (2): 1364–1406. doi:10.1093/jeea/jvaa036.

- Diamond, John B., and James P. Spillane. 2004. "High-Stakes Accountability in Urban Elementary Schools: Challenging or Reproducing Inequality." *Teachers College Record* 106 (6): 1145–1176. doi:10.1111/j.1467-9620.2004.00375.x.
- EVS. 2011. "European Values Study 2008: Integrated Dataset (EVS 2008)." ZA4800 Data file version 3.0.0. Cologne: GESIS Data Archive. doi:10.4232/1.11004.
- EVS. 2015. "Integrated Values Surveys 1981–2014" European Values Study. Accessed 5 March 2015 2017. <http://www.europeanvaluesstudy.eu>
- EVS. 2016. "EVS 2008 Method Report." GESIS Papers 2016/18. Cologne: GESIS Data Archive.
- Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer. 2016. "Penn World Table 9.0." Groningen Growth and Development Centre. doi:10.15141/S5J01T.
- Feniger, Yariv, and Adam Lefstein. 2014. "How Not to Reason with PISA Data: An Ironic Investigation." *Journal of Education Policy* 29 (6): 845–855. doi:10.1080/02680939.2014.892156.
- Fischman, Gustavo E., Amelia Marcetti Topper, Iveta Silova, Janna Goebel, and Jessica L. Holloway. 2019. "Examining the Influence of International Large-Scale Assessments on National Education Policies." *Journal of Education Policy* 34 (4): 470–499. doi:10.1080/02680939.2018.1460493.
- Fryer, Roland G, and Steven D Levitt. 2010. "An Empirical Analysis of the Gender Gap in Mathematics." *American Economic Journal. Applied Economics* 2 (2): 210–240. doi:10.1257/app.2.2.210.
- Gelfand, Michele J., Beng-Chong Lim, and Jana L. Raver. 2004. "Culture and Accountability in Organizations: Variations in Forms of Social Control across Cultures." *Human Resource Management Review* 14 (1): 135–160. doi:10.1016/j.hrmr.2004.02.007.
- Green, Andy, Jan Germen Janmaat, and Christine Han. 2009. "Regimes of Social Cohesion" *LLAKES Research Paper 1*. London: Centre for Learning and Life Chances in Knowledge Economies and Societies. Accessed 16 August 2019. <https://dera.ioe.ac.uk/10486/1/Z.-Regimes-of-Social-Cohesion.pdf>
- Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales. 2008. "Culture, Gender, and Math." *Science* 320 (5880): 1164–1165. doi:10.1126/science.1154094.
- Han, Seong Won, Francesca Borgonovi, and Sonia Guerriero. 2018. "What Motivates High School Students to Want to Be Teachers? The Role of Salary, Working Conditions, and Societal Evaluations about Occupations in a Comparative Perspective." *American Educational Research Journal* 55 (1): 3–39. doi:10.3102/0002831217729875.
- Hanushek, Eric A., and Ludger Woessmann. 2011. "The Economics of International Differences in Educational Achievement." In *Handbook of the Economics of Education*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 89–200. Vol. 3. San Diego: Elsevier. doi:10.1016/B978-0-444-53429-3.00002-8.
- Hanushek, Eric A., and Steven G. Rivkin. 2006. "Chapter 18 Teacher Quality." In *Handbook of the Economics of Education*. Vol. 2. edited by E. Hanushek, and F. Welch, 1051–1078. Amsterdam: Elsevier. doi:10.1016/S1574-0692(06)02018-6.
- Hanushek, Eric A., Susanne Link, and Ludger Woessmann. 2013. "Does School Autonomy Make Sense Everywhere? Panel Estimates from PISA." *Journal of Development Economics* 104 (September): 212–232. doi:10.1016/j.jdeveco.2012.08.002.
- Hanushek, Eric A. 2019. "Testing, Accountability, and the American Economy." *The ANNALS of the American Academy of Political and Social Science* 683 (1): 110–128. doi:10.1177/0002716219841299.
- Hanushek, Eric A, Lavinia Kinne, Philipp Lergetporer, and Ludger Woessmann. 2020. "Culture and Student Achievement: The Intertwined Roles of Patience and Risk-Taking." *Working Paper 27484. Working Paper Series*. National Bureau of Economic Research. doi:10.3386/w27484.
- He, Jia, Fons J. R. van de Vijver, and Alena Kulikova. 2017. "Country-Level Correlates of Educational Achievement: Evidence from Large-Scale Surveys." *Educational Research and Evaluation* 23 (5–6): 163–179. doi:10.1080/13803611.2017.1455288.
- Heller-Sahlgren, Gabriel. 2018. "Smart but Unhappy: Independent-School Competition and the Wellbeing-Efficiency Trade-off in Education." *Economics of Education Review* 62 (February): 66–81. doi:10.1016/j.econedurev.2017.10.005.

- Hofstede, Geert. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations across Nations*. Thousand Oaks, California: SAGE.
- Holloway, Jessica, and Jory Brass. 2018. "Making Accountable Teachers: The Terrors and Pleasures of Performativity." *Journal of Education Policy* 33 (3): 361–382. doi:10.1080/02680939.2017.1372636.
- Honig, Dan, and Lant Pritchett. 2019. "The Limits of Accounting-Based Accountability in Education (And Far Beyond): Why More Accounting Will Rarely Solve Accountability Problems." *Research on Improving Systems of Education (RISE)*. doi:10.35489/BSG-RISE-WP\_2019/030.
- Hopmann, S.T. 2008. "No Child, No School, No State Left Behind: Schooling in the Age of Accountability." *Journal of Curriculum Studies* 40 (4): 417–456. doi:10.1080/00220270801989818.
- Hwa, Yue-Yi. 2019. "Teacher Accountability Policy and Sociocultural Context: A Cross-Country Study Focusing on Finland and Singapore." Doctoral thesis., University of Cambridge. doi:10.17863/CAM.55349.
- Hwa, Yue-Yi. 2021. "Contrasting Approaches, Comparable Efficacy? How Macro-Level Trust Influences Teacher Accountability in Finland and Singapore." In *Trust, Accountability and Capacity in Education System Reform: Global Perspectives in Comparative Education*, edited by Melanie Ehren and Jacqueline Baxter, 222–251. Abingdon, Oxon; New York: Routledge. doi:10.4324/9780429344855-11.
- Ingersoll, Richard M., Lisa Merrill, and Henry May. 2016. "Do Accountability Policies Push Teachers Out?." *Educational Leadership* 73 (8): 44–49.
- Inglehart, Ronald, C., Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen, eds. 2014a. *World Values Survey: Round Five - Country-Pooled Datafile Version*. Madrid: JD Systems Institute. [www.worldvaluessurvey.org/WVSDocumentationWV5.jsp](http://www.worldvaluessurvey.org/WVSDocumentationWV5.jsp) Accessed 28 November 2018.
- Inglehart, Ronald, C., Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen, eds. 2014b. *World Values Survey: Round Six - Country-Pooled Datafile Version*. Madrid: JD Systems Institute. [www.worldvaluessurvey.org/WVSDocumentationWV6.jsp](http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp) Accessed 28 November 2018.
- Inglehart, Ronald, and Christian Welzel. 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge, UK: Cambridge University Press.
- Iyengar, Radhika. 2012. "Social Capital as the Catalyst for School Participation." *Compare* 42 (6): 839–862. doi:10.1080/03057925.2012.657930.
- Jerrim, John. 2015. "Why Do East Asian Children Perform so Well in PISA? An Investigation of Western-Born Children of East Asian Descent." *Oxford Review of Education* 41 (3): 310–333. doi:10.1080/03054985.2015.1028525.
- Jerrim, John, Luis Alejandro Lopez-Agudo, Oscar D. Marcenaro-Gutierrez, and Nikki Shure. 2017. "What Happens When Econometrics and Psychometrics Collide? An Example Using the PISA Data." *Economics of Education Review* 61 (December): 51–58. doi:10.1016/j.econedurev.2017.09.007.
- Kim, Dae Jung. 1994. "Is Culture Destiny? The Myth of Asia's Anti-Democratic Values." *Foreign Affairs* 1 (November): 1994. Accessed 22 August 2017. <https://www.foreignaffairs.com/articles/southeast-asia/1994-11-01/culture-destiny-myth-asias-anti-democratic-values>
- Koppell, Jonathan GS. 2005. "Pathologies of Accountability: ICANN and the Challenge of "Multiple Accountabilities Disorder"." *Public Administration Review* 65 (1): 94–108. doi:10.1111/j.1540-6210.2005.00434.x.
- Laukaityte, Inga, and Marie Wiberg. 2017. "Using Plausible Values in Secondary Analysis in Large-Scale Assessments." *Communications in Statistics - Theory and Methods* 46 (22): 11341–11357. doi:10.1080/03610926.2016.1267764.
- Liew, Warren Mark. 2012. "Perform or Else: The Performative Enhancement of Teacher Professionalism." *Asia Pacific Journal of Education* 32 (3): 285–303. doi:10.1080/02188791.2012.711297.
- Markus, Hazel Rose, and Alana Conner. 2013. *Clash! How to Thrive in a Multicultural World*. New York: Penguin Publishing Group.

- Markus, Hazel Rose, and Shinobu Kitayama. 2010. "Cultures and Selves: A Cycle of Mutual Constitution." *Perspectives on Psychological Science* 5 (4): 420–430. doi:10.1177/1745691610375557.
- Martin, Michael O., Ina V.S. Mullis, and Ian Hooper, eds. 2016. *Methods and Procedures in TIMSS 2015*. Chestnut Hill, Massachusetts: TIMSS & PIRLS International Study Center, Boston College.
- Martin, Michael O., Ina V.S. Mullis, Pierre Foy, and Martin Hooper. 2016. *TIMSS 2015 International Results in Mathematics*. Chestnut Hill, Massachusetts: TIMSS & PIRLS International Study Center, Boston College.
- Maxwell, Joseph A. 2012. *A Realist Approach for Qualitative Research*. Los Angeles: SAGE Publications.
- Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania." *The Quarterly Journal of Economics* 134 (3): 1627–1673. doi:10.1093/qje/qjz010.
- McDonnell, Lorraine M., and Richard F. Elmore. 1987. "Getting the Job Done: Alternative Policy Instruments." *Educational Evaluation and Policy Analysis* 9 (2): 133–152. doi:10.3102/01623737009002133.
- McLaughlin, Milbrey Wallin. 1987. "Learning From Experience: Lessons From Policy Implementation." *Educational Evaluation and Policy Analysis* 9 (2): 171–178. doi:10.3102/01623737009002171.
- Meyer, Heinz-Dieter, and Kathryn Schiller. 2013. "Gauging the Role of Non-Educational Effects in Large-Scale Assessments: Socio-Economics, Culture and PISA Outcomes." In *PISA, Power, and Policy: The Emergence of Global Educational Governance*, edited by Heinz-Dieter Meyer and Aaron Benavot, 207–224. Oxford: Symposium Books.
- Mizel, O. 2009. "Accountability in Arab Bedouin Schools in Israel: Accountable to Whom?." *Educational Management Administration and Leadership* 37 (5): 624–644. doi:10.1177/1741143209339654.
- Monaghan, Christine, and Elisabeth King. 2018. "How Theories of Change Can Improve Education Programming and Evaluation in Conflict-Affected Contexts." *Comparative Education Review* 62 (3): 365–384. doi:10.1086/698405.
- Mourshed, Mona, Marc Krawitz, and Emma Dorn. 2017. *How to Improve Student Educational Outcomes: New Insights from Data Analytics*. Discussion paper. McKinsey & Company. Accessed 3 December 2018. <https://www.mckinsey.com/~media/McKinsey/Industries/Social%20Sector/Our%20Insights/How%20to%20improve%20student%20educational%20outcomes/How-to-improve-student-educational-outcomes-New-insights-from-data-analytics.ashx>
- Müller, J., and F. Hernández. 2010. "On the Geography of Accountability: Comparative Analysis of Teachers' Experiences across Seven European Countries." *Journal of Educational Change* 11 (4): 307–322. doi:10.1007/s10833-009-9126-x.
- Muller, Jerry Z. 2018. *The Tyranny of Metrics*. Princeton: Princeton University Press.
- Murnane, Richard J., and Barbara R. Phillips. 1981. "Learning by Doing, Vintage, and Selection: Three Pieces of the Puzzle Relating Teaching Experience and Teaching Performance." *Economics of Education Review* 1 (4): 453–465. doi:10.1016/0272-7757(81)90015-7.
- Narwana, K. 2015. "A Global Approach to School Education and Local Reality: A Case Study of Community Participation in Haryana, India." *Policy Futures in Education* 13 (2): 219–233. doi:10.1177/1478210314568242.
- Norris, Pippa, and Ronald Inglehart. 2004. *Sacred and Secular: Religion and Politics Worldwide*. Cambridge, UK: Cambridge University Press.
- OECD. 2009. *PISA Data Analysis Manual: SPSS Second Edition*. Paris: OECD Publishing.
- OECD. 2014a. *PISA 2012 Technical Report*. Paris: OECD Publishing.
- OECD. 2014b. *PISA 2012 Results: What Students Know and Can Do (Volume I, Revised Edition)*. Paris: Organisation for Economic Co-operation and Development.
- OECD. 2016a. *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing.

- OECD. 2016b. "PISA 2015 Annex B1.4 Results (Tables): Governance, Assessment and Accountability." Dataset. OECD Code. Accessed 2 September 2019. <http://statlinks.oecdcode.org/982016071p1t004.xlsx>
- OECD. 2017a. *PISA 2015 Technical Report*. Paris: OECD Publishing.
- OECD. 2017b. "School Questionnaire for PISA 2018: Main Survey Version" OECD Publishing. Accessed 19 February 2021. [http://www.oecd.org/pisa/data/2018database/CY7\\_201710\\_QST\\_MS\\_SCQ\\_NoNotes\\_final.pdf](http://www.oecd.org/pisa/data/2018database/CY7_201710_QST_MS_SCQ_NoNotes_final.pdf)
- OECD. 2020. *PISA 2018 Results (Volume V): Effective Policies, Successful Schools*. PISA. Paris: PISA, OECD Publishing. doi:10.1787/ca768d40-en.
- Pawson, Ray, and Nick Tilley. 1997. *Realistic Evaluation*. London: SAGE Publications.
- Pollitt, Christopher, and Geert Bouckaert. 2017. *Public Management Reform: A Comparative Analysis - into the Age of Austerity*. 4th ed. New York, NY: Oxford University Press.
- Pritchett, Lant, and Martina Viarengo. 2021. "Learning Outcomes in Developing Countries: Four Hard Lessons from PISA-D." *RISE Working Paper*. 21/069: Research on Improving Systems of Education (RISE). doi:10.35489/BSG-RISE-WP\_2021/069.
- Rizvi, Fazal, and Bob Lingard. 2009. *Globalizing Education Policy*. Abingdon, Oxon; New York: Routledge.
- Rodríguez-Planas, Nùria, and Natalia Nollenberger. 2018. "Let the Girls Learn! It Is Not Only about Math . . . It's about Gender Social Norms." *Economics of Education Review* 62 (February): 230–253. doi:10.1016/j.econedurev.2017.11.006.
- Romzek, Barbara S., and Melvin J. Dubnick. 1987. "Accountability in the Public Sector: Lessons from the Challenger Tragedy." *Public Administration Review* 47 (3): 227–238. doi:10.2307/975901.
- Rubin, Donald B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91 (434): 473–489. doi:10.2307/2291635.
- Rutkowski, Leslie, Eugenio Gonzalez, Marc Joncas, and Matthias Von Davier. 2010. "International Large-Scale Assessment Data: Issues in Secondary Analysis and Reporting." *Educational Researcher* 39 (2): 142–151. doi:10.3102/0013189X10363170.
- Schleicher, Andreas. 2018. *World Class: How to Build a 21st-Century School System* Paris: OECD.
- Sellar, Sam, and Bob Lingard. 2013. "The OECD and Global Governance in Education." *Journal of Education Policy* 28 (5): 710–725. doi:10.1080/02680939.2013.779791.
- Sen, Amartya Kumar. 1999. "Democracy as a Universal Value." *Journal of Democracy* 10 (3): 3–17. doi:10.1353/jod.1999.0055.
- Spillane, James P. 2009. *Standards Deviation: How Schools Misunderstand Education Policy*. Cambridge, Massachusetts; London: Harvard University Press.
- Steiner-Khamsi, Gita. 2014. "Cross-National Policy Borrowing: Understanding Reception and Translation." *Asia Pacific Journal of Education* 34 (2): 153–167. doi:10.1080/02188791.2013.875649.
- Stout, Lynn. 2010. *Cultivating Conscience: How Good Laws Make Good People*. Princeton: Princeton University Press.
- Survey Research Center. 2016. *Guidelines for Best Practice in Cross-Cultural Surveys*. 4th ed. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.
- Taras, Vas, Bradley L. Kirkman, and Piers Steel. 2010. "Examining the Impact of Culture's Consequences: A Three-Decade, Multilevel, Meta-Analytic Review of Hofstede's Cultural Value Dimensions." *Journal of Applied Psychology* 95 (3): 405–439. doi:10.1037/a0018938.
- Thiel, Corrie, Sebastian Schweizer, and Johannes Bellmann. 2017. "Rethinking Side Effects of Accountability in Education: Insights from a Multiple Methods Study in Four German School Systems." *Education Policy Analysis Archives* 25 (August): 93. doi:10.14507/epaa.25.2662.
- Thompson, Michael, Richard Ellis, and Aaron Wildavsky. 1990. *Cultural Theory*. Boulder, Colo: Westview Press.
- Tulowitzki, Pierre. 2016. "Educational Accountability around the Globe: Challenges and Possibilities for School Leadership." In *Educational Accountability: International Perspectives on Challenges and Possibilities for School Leadership*, edited by Jacob II Easley, and Pierre Tulowitzki. London; New York: Routledge 233–238 .
- UNESCO. 2017. *Global Education Monitoring Report 2017/2018: Accountability in Education – Meeting Our Commitments*. Paris: UNESCO.



- Velayutham, S., and M. H. B. Perera. 2004. "The Influence of Emotions and Culture on Accountability and Governance." *Corporate Governance: The International Journal of Business in Society* 4 (1): 52–64. doi:10.1108/14720700410521961.
- Verger, Antoni, and Lluís Parcerisa. 2017. "A Difficult Relationship. Accountability Policies and Teachers: International Evidence and Key Premises for Future Research Akiba, Motoko, and LeTendre, Gerald K. eds." In *International Handbook of Teacher Quality and Policy*, 241–254. New York: Routledge. doi:10.5281/zenodo.1256602.
- Von der Embse, Nathaniel P., Laura L. Pendergast, Natasha Segool, Elina Saeki, and Shannon Ryan. 2016. "The Influence of Test-Based Accountability Policies on School Climate and Teacher Stress across Four States." *Teaching and Teacher Education* 59 (October): 492–502. doi:10.1016/j.tate.2016.07.013.
- Wagemaker, Hans. 2010. "IEA: Globalization and Assessment". In *International Encyclopedia of Education* (Oxford: Academic Press), edited by Penelope Peterson, Eva Baker, and Barry McGaw, 663–668. doi:10.1016/B978-0-08-044894-7.01477-9.
- Wagner, Robert B. 1989. *Accountability in Education: A Philosophical Inquiry*. New York: Routledge.
- Webber, D.J. 2010. "School District Democracy: School Board Voting and School Performance." *Politics and Policy* 38 (1): 81–95. doi:10.1111/j.1747-1346.2009.00229.x.
- West, Martin R., and Ludger Woessmann. 2010. "“Every Catholic Child in a Catholic School”: Historical Resistance to State Schooling, Contemporary Private Competition and Student Achievement across Countries." *The Economic Journal* 120 (546): F229–55. doi:10.1111/j.1468-0297.2010.02375.x.
- Williams, Martin J. 2020. "External Validity and Policy Adaptation: From Impact Evaluation to Policy Design." *The World Bank Research Observer* 35 (2): 158–191. doi:10.1093/wbro/lky010.
- Woessmann, Ludger. 2016. "The Importance of School Systems: Evidence from International Differences in Student Achievement." *Journal of Economic Perspectives* 30 (3): 3–32. doi:10.1257/jep.30.3.3.
- Woessmann, Ludger, Elke Luedemann, Gabriela Schuetz, and Martin R. West. 2009. *School Accountability, Autonomy and Choice around the World*. Cheltenham; Northampton: Massachusetts: Edward Elgar Publishing .
- WVS Association. n.d. "WVS Database: Fieldwork and Sampling" *Official web site*. World Values Survey. Accessed 11 August 2017. <http://www.worldvaluessurvey.org/WVSContents.jsp>
- You, Yun. 2017. "Comparing School Accountability in England and Its East Asian Sources of “Borrowing”." *Comparative Education* 53 (2): 224–244. doi:10.1080/03050068.2017.1294652.
- Zhao, Yong. 2018. *What Works May Hurt: Side Effects in Education*. New York: Teachers College Press.